UNIVERSIDAD CARLOS III DE MADRID
Departamento de Ingeniería Telemática

Doctorado en Tecnologías de las Comunicaciones

Tesis Doctoral

CONTRIBUCION AL DISEÑO DE REDES CAMPUS ETHERNET AUTOCONFIGURABLES

Autor: Guillermo Ibáñez Fernández
Ingeniero de Telecomunicación

Director: Arturo Azcorra Saloña
Doctor Ingeniero de Telecomunicación

Octubre de 2005

Guillermo A. Ibáñez Fernández
Departamento de Ingeniería Telemática.
Universidad Carlos III Madrid.
Avenida de la Universidad, 30.
28911 Leganés (Madrid).
e-mail: gibanez@it.uc3m.es

Título: Contribución al diseño de redes campus Ethernet autoconfigurables
Autor: Guillermo Ibáñez Fernández
Director: Dr. D. Arturo Azcorra Saloña

El Tribunal nombrado por el Magnífico y Excelentísimo Sr. Rector de l Universidad Carlos III de Madrid el día ____ de _____de _____.

Presidente: _____

Vocal: _____

Vocal: _____

Vocal: _____

Secretario: _____

Realizado el acto de defensa y lectura de la Tesis el día
en en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

LOS VOCALES

EL SECRETARIO EL PRESIDENTE

A mis padres.
A mi esposa.
A mis hijos.

Este trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia Español a través del proyecto CAPITAL (TEC2004-05622-C04-03/TCM).

Agradecimientos

Deseo expresar mi agradecimiento a todos los que han hecho posible la elaboración de esta Tesis.

En primer lugar a mi director de tesis, Arturo Azcorra, quien me ha aportado su gran experiencia, claridad de visión y su continuo apoyo, a pesar de sus múltiples responsabilidades y proyectos; a mi tutor de doctorado Alberto García por la generosidad con su tiempo, rigor y sabios consejos.

A mis profesores de doctorado y compañeros Ignacio Soto, David Larrabeiti, María Calderón, Jaime García, Francisco Valera, José Félix Kukielka, José Ignacio Moreno y tantos otros, por lo que he aprendido de ellos y de su ejemplo profesional, por su estímulo, consejos y opiniones sinceras.

A todos mis compañeros doctorandos de la Universidad Carlos III, con quienes he recorrido este interesante camino, que es individual y colectivo a la vez, con muy estimulantes discusiones. En particular quiero agradecer a Marcelo Bagnulo y a Carlos Jesús Bernardos por mantenerme informado y a Iván Vidal y Ricardo Romeral por su ayuda y compañerismo.

Merecen una mención especial Mar Yébenes por la ayuda con las simulaciones y Matt Hutton por sus revisiones de texto en inglés.

También agradezco a mis queridos amigos y compañeros Carlos López y Belén Ruiz de la Universidad Carlos III su orientación y ayuda.

Quiero expresar un agradecimiento especial a Francisco González Vidal, por su orientación, ánimo y apoyo en los primeros pasos de este camino y con quien he tenido el privilegio de trabajar en proyectos promovidos por él en Alcatel.

Y en general a todos los que me ayudaron, inspiraron o alentaron en este esfuerzo.

Finalmente, a mi hermana Ana por animarme a hacerlo, a mi esposa Enriqueta y a mis hijos, que han soportado con paciencia mis presencias “ausentes” en este período.

Resumen

Los conmutadores Ethernet tienen muy altas prestaciones, coste moderado y configuración mínima, lo que los hace muy adecuados para redes campus, pero los protocolos de capa dos actuales no tienen la escalabilidad suficiente para utilizarse en redes campus de tamaño medio y precisan encaminadores (routers) que compartimenten la red. Estas redes campus, de creciente tamaño, capacidad y complejidad, requieren nuevos dispositivos que superen las limitaciones de escalabilidad de los switches y la complejidad de configuración de los routers.

Este documento propone una arquitectura y varios protocolos y dispositivos, (denominados como Conmutadores de Funcionalidad Añadida), para redes campus Ethernet de gran tamaño autoconfigurables, que constituye una solución completa, escalable y adaptable. La arquitectura propuesta consta de dos capas o niveles jerárquicos: núcleo y acceso. El núcleo o troncal está formado por Bridges de Camino Mínimo que encaminan a través de árboles múltiples de expansión enraizados en cada bridge frontera. La capa de acceso está formada por bridges estándar conectados a un bridge frontera del núcleo. Cada bridge del núcleo a su vez actúa como raíz del árbol de expansión de la red de acceso conectada al mismo. Puede opcionalmente optimizarse el ancho de banda y la utilización de los enlaces en la capa de acceso con la nueva funcionalidad propuesta de Routing Bridges. Estos bridges utilizan caminos redundantes más cortos entre bridges, por enlaces que normalmente son bloqueados por el protocolo de árbol de expansión. Ambos tipos de Conmutadores de Funcionalidad Añadida constituyen una evolución de los bridges transparentes, combinando las ventajas de los routers (camino mínimo, rápida convergencia, utilización eficiente de la red), con la sencillez de configuración de los bridges. Se desarrollan asimismo propuestas para la limitación del tráfico multicast y broadcast en la red campus. Se han contrastado mediante simulaciones y estudios analíticos las prestaciones de la arquitectura y el bajo nivel de configuración requerido. La reciente propuesta técnica de Shortest Path Bridging en el IEEE, con principios comunes aunque diferente en su planteamiento y requisitos, parece confirmar la adecuación del enfoque utilizado en la Tesis basado en árboles múltiples de expansión autoenraizados para el diseño de redes campus y metropolitanas.

Palabras clave: redes de computadoras, protocolos, árboles múltiples de expansión.

Abstract

Current Ethernet switches features like high performance, low cost and zero configuration features make them suitable for big campus networks, but standard layer two protocols do not scale to medium size campus networks without routers to segment the switched domain. These networks, with its increasing size, bandwidth and complexity, require new devices that overcome the limitations of switches and the complexity of configuration of routers.

This paper proposes architecture and several protocols and devices for self configuring big size Ethernet campus networks that is a complete, flexible and scalable solution. The devices can be classified as Added Functionality Switches. The architecture consists of two hierarchical layers: Core and Access layer. The Core layer is formed by bridges that perform bridging via shortest paths using multiple spanning trees rooted at edge bridges. Standard bridges connected to the Core bridges constitute the Access Layer. Every Core bridge acts as root of the spanning tree of Access Layer network connected to it. Performance of Access Layer can be optionally enhanced using a new type of Routing Bridge that uses shorter redundant paths between bridges, normally blocked by the standard spanning tree protocol. These two types of switches are hybrid bridges that can be classified as Added Functionality Switches, more specifically Shortest Path Bridges or Routing Bridges. These bridges are an evolution of Transparent Bridges, hybrid devices that combine the advantages of routers (minimum path, rapid convergence, full infrastructure utilization), with the configuration simplicity of bridges. Proposals for limitation of multicast and broadcast traffic in the switched network are also described. The simulations and analysis show the high performance and low configuration required of the proposed architecture. The recent IEEE preliminary technical paper for Shortest Path Bridging, sharing some principles but different in requirements and approach, seems to confirm the suitability of the self rooting multiple spanning trees approach proposed in this These to campus and metropolitan networks design.

Keywords: protocols, computer networks, multiple spanning tree, shortest path bridging.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y objetivos

1.1 Introducción

En este capítulo se presenta la situación de las redes campus Ethernet dentro de su contexto actual y su evolución anterior. Para ello se describen los cambios habidos, graduales pero esenciales, en la evolución reciente de Ethernet haciendo énfasis en las principales limitaciones que presentan dichas redes. Se adelanta una primera definición del problema a resolver y de los requisitos a cumplir. Finalmente se describe la organización de la Tesis.

Actualmente, las prestaciones y economía de Gigabit Ethernet y 10 GE empujan hacia cambios importantes tanto en las redes campus como en las redes metropolitanas (MAN), cambios provocados por los drásticos incrementos en ancho de banda y escalabilidad así como por la facilidad de gestión y el bajo coste respecto a los conmutadores ATM y a los equipos de Jerarquía Digital Síncrona (SDH). Dichos avances, junto con la creciente demanda de ancho de banda por las aplicaciones multimedia (formación multimedia on-line, videoconferencia, telecongresos, etc.) y el creciente número de equipos conectados a las redes campus plantean la necesidad de disponer de redes campus Ethernet autoconfigurables, de altas prestaciones, escalables a grandes tamaños de red y de coste reducido.

1.2 Evolución de Ethernet

Ethernet se ha consolidado en los últimos años como el estándar predominante en las redes locales de ámbito empresarial y metropolitano frente a FDDI, ATM, SDH, DQDB y otros, por sus altas prestaciones, compatibilidad con equipos Ethernet de distintas velocidades, economía, capacidad de autoconfiguración e independencia del direccionamiento IP. Ethernet, inicialmente a 3 Mbps, ha evolucionado en capacidad desde 10 Mbps a 10 Gb en 23 años y de los simples bridges que unían dos redes locales se ha pasado a los conmutadores de N*10 Gigabit**.** Su evolución se acerca a la conocida "ley" de Moore de duplicación de la capacidad de los dispositivos semiconductores cada 24 meses [Moo65].

Esta evolución está planteando cambios importantes en la arquitectura de las redes LAN, MAN y WAN. Las redes campus están migrando a Ethernet. Las redes campus basadas en Ethernet, al hacerse cada vez más extensas, veloces y complejas, requieren nuevas funcionalidades en los nodos que garanticen escalabilidad, rendimiento y seguridad requiriendo a la vez mínima gestión y configuración. Los retos principales en redes campus derivan de su naturaleza conmutada (tormentas de tramas), de la compartición del medio (rendimiento y seguridad), o de su complejidad de configuración (árboles múltiples de expansión por grupos de VLAN, gestión y configuración de VLANs) y de una configuración y administración de direcciones IP consistente con las VLAN configuradas.

Las redes Ethernet han evolucionado desde utilizar en sus orígenes un medio compartido a ser punto a punto y desde rango LAN a distancias de rango WAN. Este ha sido un cambio gradual, pero radical desde el punto de vista de Ethernet, porque supone la práctica desaparición del medio compartido, algo esencial en los orígenes de Ethernet. Anteriormente en las redes de area amplia WAN se utilizaba conmutación mientras que en el ámbito local (redes LAN), las tecnologías se basaban en compartir el medio. Las redes locales en general, y Ethernet en particular, nacieron siendo esencialmente de medio compartido (*the Ether*) de alta capacidad, frente a las tecnologías WAN que se basaban en conmutación, y con caudales de transmisión inferiores en general a los disponibles en las LAN. Sin embargo, la evolución de las tecnologías LAN y WAN está haciendo converger las soluciones tecnológicas utilizadas en ambos entornos. Las redes LAN basadas en Ethernet de la actualidad son mayoritariamente conmutadas, punto a punto, full-duplex, incorporan multiplexación mediante IEEE 802.1Q (VLAN) y soportan distancias de transmisión idénticas a las soportadas por los enlaces WAN convencionales. Entendemos que el desarrollo de los conmutadores transparentes Ethernet ha hecho posible esta evolución.

La tendencia hacia redes Ethernet sin medio compartido es confirmada por los estándares del IEEE 802.1X, 802.1w (RSTP, ahora en 802.1D), 802.1Q (MSTP) los cuales establecen como requisito los enlaces punto a punto. Los enlaces dedicados son necesarios para obtener las máximas prestaciones, hacer efectiva la seguridad en el nivel de enlace y poder simplificar los protocolos de enlace. Un enlace Ethernet compartido es mucho menos eficiente que uno dedicado debido a las colisiones. Los enlaces dedicados hacen posibles mecanismos de convergencia rápida en capa dos como el empleado por RSTP. Como ejemplo, el estándar 10 GE no contempla ni siquiera el uso de enlaces half duplex como lo hacía Gigabit Ethernet donde se mantuvo por razones de compatibilidad.

En la Ethernet original predominaba la difusión e inundación como mecanismos básicos y hoy se busca la mínima difusión de las tramas por las mismas razones que se evitan en las redes WAN: rendimiento y control de tráfico.

Paralelamente, las redes WAN van incorporando las tecnologías LAN por su robustez y su buena relación precio/prestaciones, estando ya bastante implantadas en entornos de acceso y metropolitanos, y de forma creciente en el entorno propiamente de área extensa. Esta tendencia continuará en el futuro a medida que las redes campus aumentan de tamaño, capacidad y complejidad. Las funcionalidades actuales de las redes campus incluyen agregación de enlaces para fiabilidad, reparto de carga y posibilidad de ampliación gradual de la red, redes locales virtuales (VLAN), árbol de distribución múltiple (por VLAN), inspección de IGMP en capa 2 y algunas otras, que hacen más crítica y compleja su gestión.

Otra tendencia complementaria, a la que se presta atención en esta Tesis y que consideramos relevante es que las redes campus, al aumentar significativamente su tamaño, tenderán a parecerse en sus funciones y soluciones las redes metropolitanas, si bien manteniendo importantes diferencias derivadas de su distinta explotación y propiedad, (las redes metropolitanas propiedad de operadores de red y orientadas a dar servicios de conectividad a múltiples usuarios y las redes campus propiedad de las empresas usuarias finales).

Desde el punto de vista de funciones, el desarrollo tecnológico de los conmutadores ha incorporado diversas funcionalidades a la simple conmutación. Desde el punto de vista del estándar, el formato de la trama Ethernet estándar se ha extendido respecto a la inicial, habiéndose incorporado el etiquetado VLAN (802.1Q) y el de prioridades para establecer clases de servicio (802.1p). Las actividades en los grupos de trabajo del IEEE respecto a la ampliación de la trama y definición de nuevas etiquetas han aumentado en los últimos años. Estas actividades tienen como objetivo extender las funcionalidades de Ethernet tales como introducir cifrado en capa 2, pero principalmente separar tráfico en redes de proveedores de servicios o de ámbito metropolitano como se explica para Redes Metropolitanas Ethernet y Redes Privadas Virtuales de capa 2 (L2VPN) en el capítulo 3.

1.3 Limitaciones del Protocolo de Árbol de Expansión

El mecanismo estándar hasta 2004 para difusión en capa 2 en las redes Ethernet era el Protocolo de Árbol de Expansión (AE, Spanning Tree Protocol)[STP]. El protocolo de Árbol de Expansión construye un árbol de distribución de tramas que corta algunos enlaces eliminándolos de la topología activa para evitar bucles. El AE minimiza la distancia desde el bridge raíz hasta cada nodo de la red, pero las distancias entre nodos a través del AE no son mínimas. Adicionalmente, el tráfico se distribuye atravesando el bridge raíz, lo que crea cuellos de botella en sus proximidades. STP está basado en temporizadores y tiene tiempos de convergencia superiores a 30 segundos, lo que es difícilmente aceptable en entornos metropolitanos como los que actualmente utilizan

Jerarquía Digital Síncrona, que disponen de mecanismos de reconfiguración rápida en 50 milisegundos.

Hace unos años fue estandarizado el Protocolo Rápido de Árbol de Expansión [RSTP]. Este protocolo reduce el tiempo de convergencia ante cambios de topología al rango desde decenas de milisegundos hasta un segundo, haciendo uso de la circunstancia de que los enlaces Ethernet en las redes actuales son de forma predominante enlaces dedicados. Pero RSTP sigue utilizando un árbol de expansión como topología activa para la difusión, por lo que adolece de los restantes inconvenientes de STP mencionados más arriba, principalmente que los caminos no son mínimos y el tráfico se concentra alrededor del bridge raíz.

Otro estándar relativamente reciente es el Protocolo de Árbol de Expansión Múltiple (Multiple Spanning Tree Protocol)[802.1D]. MSTP está basado en la configuración de instancias múltiples simultáneas de árboles de expansión en una determinada *región.* Se asignan mediante configuración diferentes árboles, prediseñados manualmente, a las distintas VLANs, lo que permite encaminar el tráfico separadamente por redes virtualmente independientes implementadas sobre una única red física. De esta forma se distribuye mejor el tráfico y mejora notablemente la utilización de la infraestructura de red porque los árboles múltiples habilitan y utilizan enlaces que serían cortados por el protocolo (caso de utilizar solamente STP ó RSP) para evitar la formación de bucles.

Pero el uso de MSTP presenta varios inconvenientes que restringen severamente su uso. El principal inconveniente es la complejidad de diseño y configuración de los árboles múltiples, planificando sus bridges raíz respectivos y la asignación de VLANs a árboles, que debe ser configurada de forma idéntica en todos los bridges de la región MSTP. La configuración adecuada de VLANs en la red es delicada porque puede segmentarse la red en caso de fallo de un enlace al no poderse reencaminar alguna VLAN. Otro inconveniente es el diseño de dichas regiones MSTP dentro de las que operan estos árboles múltiples. Finalmente, los caminos utilizados en los árboles de expansión no son mínimos.

1.4 Definición del problema.

En esta sección se adelanta una breve definición del problema a resolver.

El problema planteado en la presente Tesis consiste en definir una arquitectura de alta capacidad adecuada para las redes campus futuras, y que a la vez permita simplificar la configuración y mantenimiento de la red, optimizando el uso de la infraestructura de comunicaciones mediante la reducción del coste de los equipos. Los tamaños de redes que se contemplan son significativamente superiores a los actuales, dada la previsible proliferación de dispositivos de todo tipo en las redes campus. Consideramos una Red campus de hasta 100.000 dispositivos de los cuales aproximadamente 20.000 pueden ser ordenadores convencionales, siendo el resto dispositivos de diversos tipos: sensores, paneles, asistentes personales (PDA), etc. El numero de bridges típico para una red de este tipo puede ser de alrededor de 500.

Actualmente las redes campus presentan los siguientes inconvenientes:

- Deben fragmentarse los dominios de conmutación para limitar la propagación de problemas tales como tormentas de tramas. Para ello deben situarse Routers o Conmutadores Multicapa para fragmentar en subredes más pequeñas.
- Es necesario configurar y administrar las redes LAN virtuales (VLANs) así como diseñar y configurar los Árboles de Expansión por VLAN para separar el tráfico y los dominios de difusión.
- Hay que asignar y gestionar las direcciones IP, y la dirección IP cambia al cambiar el el usuario de lugar de conexión.
- Se infrautiliza mucha infraestructura costosa debido a los enlaces bloqueados por el protocolo de Árbol de Expansión (STP).

En resumen, se trata de definir conmutadores Ethernet de funcionalidad añadida que mantengan las mantengan las ventajas de los bridges eliminando sus inconvenientes y sus protocolos de funcionamiento que implementen redes Ethernet lo más autoconfigurables posible.

1.5 Requisitos básicos.

En esta sección realizamos una primera aproximación a los requisitos a cumplir. Para establecer los requisitos básicos de las redes campus se parte de las carencias y problemas de las redes actuales y de la extrapolación de las tendencias actuales. En cuanto a número y tipos de dispositivos se extrapola el crecimiento previsible en número de dispositivos y ancho de banda requerido según las tecnologías y aplicaciones disponibles y predecibles actualmente. Es previsible el incremento en el número de sensores de todo tipo, PDAs, pantallas informativas de visualización, dispositivos inalámbricos, etc., conectados a redes campus. Respecto al ancho de banda, el despliegue masivo de aplicaciones multimedia (videoconferencia, formación multimedia) puede requerir en el futuro capacidades muy superiores a las actualmente instaladas, y se producirá en la medida de que su despliegue resulte económico en términos de coste-beneficio. Por otro lado es necesario determinar los requisitos específicos de red derivados del gran incremento cuantitativo en tamaño de la red y ancho de banda, o visto de otra manera, resolver los posibles problemas aparecidos al crecer el tamaño del dominio conmutado.

En cuanto a la matriz de distribución de tráfico en las redes campus, se reconoce generalmente [ZM04] la dificultad de predecir a medio plazo la matriz de distribución de tráfico de los núcleos de redes. Se parte de la situación actual de predominio absoluto del modelo cliente-servidor, por lo que el tráfico entre sistemas finales será poco relevante. Dado que la matriz de distribución de tráfico está sujeta a cambios impredecibles, como podría ser un aumento importante del tráfico Peer-to-Peer originado por el éxito de dicho modelo a nivel local en las redes campus, esto haría deseable la característica de independencia de congestiones en la red de la matriz de tráfico mediante mecanismos de reparto de carga, pero sin elevarlo a la categoría de requisito.

A continuación se establecen los requisitos principales de las redes campus futuras. Son los siguientes:

-Interoperabilidad con routers IP. La operación con los routers IP actuales debería ser similar a la actual.

-Escalabilidad. Nuestro escenario contempla redes campus de hasta 100.000 dispositivos, de los cuales alrededor de 20.000 PCs, y unos 500 bridges.

-Configuración mínima o nula. Como criterio básico relativo a costes de configuración, el esfuerzo de configuración por dispositivo debería ser inversamente proporcional al número de dispositivos de ese tipo en la red. Desde el punto de vista de costes de explotación de la red, es más importante que un dispositivo sea autoconfigurable cuanto más numeroso sea. Desde el punto de vista funcional, es más importante que un dispositivo sea autoconfigurable cuanto mayor sea el daño producido en caso de configuración errónea.

-Uso eficiente de la infraestructura de comunicaciones. El tráfico debe distribuirse en toda la infraestructura instalada. De forma lo más equilibrada posible en tanto sea compatible con los requisitos de simplicidad y economía.

-Transparencia. Los nuevos bridges introducidos deben ser transparentes para los sistemas finales y para los Routers IP y en general para todos los dispositivos.

-Compatibilidad Ethernet. La interoperabilidad con bridges estándar 802.1D es un requisito muy deseable para posibilitar una introducción progresiva en la red.

-La seguridad de esta red debe ser igual o mejor que la de los bridges actuales.

-Como requisito relativo a reconfiguración ante fallos es de destacar la reconfiguración automática en el menor tiempo posible.

-Deben obtenerse caminos más cortos que los del Árbol de Expansión.

1.6 Estructura de la Memoria

A fin de facilitar la lectura de la Tesis y dar una visión general de la misma, se describe a continuación su estructura y el contenido de los capítulos. La Tesis está estructurada de la forma siguiente: en el capítulo 1 se ha presentado el contexto actual de las redes campus Ethernet, su evolución, tendencias y carencias actuales; se ha adelantado una definición del problema a resolver y de los requisitos. En el Capitulo 2 se analiza el estado del arte en las áreas directamente relacionadas y en otras áreas colaterales, exploradas por presentar aplicabilidad potencial de conceptos. En el Capitulo 3 se plantea el problema a resolver, se establecen los requisitos y se describe el proceso de diseño. En el Capitulo 4 se realiza una descripción general de la red en dos niveles propuesta así como de los dispositivos empleados y su funcionalidad. En el Capítulo 5 se describen los protocolos propuestos para ambos niveles de la red. En el Capítulo 6 se describe la evaluación de prestaciones de la arquitectura y protocolos propuestos y las simulaciones realizadas. Finalmente en el Capítulo 7 se elaboran las principales conclusiones y se describen las orientaciones para el trabajo futuro.

Capítulo 2

Estado del arte

En este capítulo se revisa el estado del arte y los antecedentes en redes conmutadas autoconfigurables. Se organiza el capítulo agrupando el trabajo relacionado por tecnologías, por semejanza en el tipo de soluciones propuestas y dentro de las mismas por orden cronológico.

El problema planteado en esta Tesis afecta a la funcionalidad principal de los bridges, lo que ha producido numerosas aportaciones de diversa relevancia repartidas en un largo período desde la invención de los bridges transparentes. Por otra parte, los aspectos concretos de diseño de las redes campus confieren una amplitud considerable al análisis del estado del arte a la vez que obligan además a revisar la realidad práctica del diseño de redes campus para completar y validar las propuestas académicas, confiriendo un doble enfoque, teórico y práctico, al problema planteado. La exposición se ha estructurado de la forma siguiente, que se considera más idónea para su comprensión:

- Protocolos de Árbol Único de Expansión.
- Protocolos de Árboles de Expansión con caminos alternativos.
- Protocolos de Árboles Múltiples de Expansión
- Protocolos de Prohibición de Giros.
- Bridges con encaminamiento y combinaciones de bridge y router.
- Bridging jerárquico.
- Hardware:
  - Evolución de las redes campus.
  - Arquitectura de las redes campus
  - Estado del arte de conmutadores.
- Conceptos de otras áreas aplicables a la solución:
  - Direccionamiento y encaminamiento jerárquicos.

- Asignación de direcciones, autoconfiguración.
- Direccionamiento y encaminamiento en redes Peer to Peer.

- Estado del arte de optimización de tráfico multicast.

2.1 Protocolos de Árbol Único de Expansión.

En esta sección se introducen los protocolos de árbol de expansión y se describen sumariamente los protocolos que utilizan un único árbol de expansión estandarizados por el IEEE: STP [STP] y RSTP [RSTP]. Primeramente se describe STP y a continuación el protocolo que recientemente le ha sucedido en el estándar 802.1D, el Protocolo Rápido de Árbol de Expansión (RSTP).

El funcionamiento de los bridges transparentes consta de tres mecanismos básicos:

- Construcción de un árbol de expansión.
- Difusión de las tramas por el árbol de expansión.
- Aprendizaje de direcciones MAC asociadas a cada puerto.

La función del árbol de expansión es facilitar una estructura de difusión de las tramas por toda la red de forma que cada trama llegue a todos los nodos, conservándose el orden de las tramas y sin duplicación de las mismas ni formación de bucles. Esta estructura de difusión es el conocido Árbol de Expansión (*Spanning Tree*). El problema de construir un Árbol Mínimo de Expansión es un problema clásico dentro de los algoritmos centralizados destacando los algoritmos de Prim [Pri57] y Kruskal [Kru56]. El algoritmo de Prim construye el árbol de mínima longitud partiendo de un nodo añadiendo sucesivamente el nodo de menor coste de unión al árbol. Es una derivación del algoritmo de Djikstra de caminos mínimos (SPF) y se diferencia en que Djikstra incorpora en cada paso el nodo con menor de menor coste total hasta el nodo origen. El algoritmo de Kruskal va seleccionando los enlaces de menor coste de la red y forma un bosque de árboles disjuntos que se terminan uniendo en un único árbol.

La formación de un árbol mínimo de expansión de forma distribuida plantea dificultades adicionales, como la formación de bucles si no existen mecanismos de desempate tales como disponer de una identidad distinta en cada nodo [Lyn96]. En la literatura existen diversos algoritmos distribuidos de construcción de Árboles Mínimos de Expansión [GHS83]. Como se ha indicado, el protocolo estándar actual para difusión en capa 2 en las redes Ethernet es el Protocolo de Árbol Rápido de Expansión Rapid Spanning Tree Protocol (RSTP)) [RSTP].

Un Protocolo de Árbol de Expansión (PAE) como STP o RSTP construye un árbol de distribución de tramas a todos los nodos de la red cortando algunos enlaces entre nodos eliminándolos de la topología activa para evitar bucles. El PAE minimiza la distancia desde el bridge raíz hasta cada nodo de la red, pero no minimiza las distancias entre nodos a través del árbol. Adicionalmente, el tráfico se canaliza por el árbol y cambia de rama atravesando el bridge raíz, lo que crea cuellos de botella en sus proximidades. Por ello STP supone una limitación importante actualmente en las redes campus. Unido ello a los tiempos de convergencia del orden de 45 segundos de STP, hace

que en bastantes casos se evite su uso, utilizando topologías intrínsecamente libres de bucles.

El protocolo de Árbol Rápido de Expansión (Rapid Spanning Tree Protocol (RSTP)) reduce el tiempo de convergencia tras cambios en la topología a un rango de entre decenas de milisegundos y un segundo, haciendo uso de la circunstancia de que los enlaces Ethernet en las redes actuales son enlaces predominantemente dedicados. Pero RSTP sigue utilizando un árbol de expansión como topología activa para la difusión, por lo que adolece de los restantes inconvenientes de STP mencionados más arriba, principalmente que los caminos no son mínimos y que el tráfico se concentra alrededor del bridge raíz.

2.1.1 Protocolo de Árbol de Expansión (STP)

El protocolo STP se basa en el algoritmo creado por Radia Perlman [Per85] y fue incorporado al estándar IEEE 802.1D. El algoritmo minimiza el coste de cada nodo al nodo elegido como raíz. Se basa en el intercambio de unidades de protocolo de bridge (BPDUs) entre los bridges de una red. Cada bridge envía BPDUs a sus vecinos tras recibir una BPDU procedente del bridge raíz. El proceso de formación del árbol de expansión en una red consta de tres pasos: Elegir el bridge raíz, elegir el puerto raíz en cada bridge (puerto que conecta al árbol hasta el bridge raíz) y elegir los puertos designados (puertos que prolongarán el árbol hacia abajo a otros bridges). La identidad de bridge *(bridge ID)* se compone de un prefijo de prioridad configurable de 16 bit y un sufijo de 48 bit que contiene la dirección MAC del bridge. Los bridges eligen como bridge raíz el de menor identidad de bridge. En este proceso todos los bridges comienzan autoreivindicándose como bridge raíz mientras no reciben BPDUs de otros bridges con bridge ID menor que la suya. Tras un tiempo de intercambio de BPDUs entre los bridges, todos los bridges coinciden en la elección del bridge con menor bridge ID de la red. La elección de puerto raíz se hace con los criterios siguientes: cada bridge elige el de menor coste de camino hasta el bridge raíz y los posibles empates se resuelven eligiendo el de menor identidad de bridge designado de la BPDU y si hubiera empate, el de menor identidad de puerto designado.

El procesado de las BPDUs es como sigue: el bridge raíz envía periódicamente (cada Hello Time, dos segundos por defecto) BPDUs de configuración del protocolo STP a los bridges vecinos. Estas BPDUs contienen la identidad del bridge raíz, la identidad del propio bridge, el coste del camino hasta el bridge raíz y la identidad del puerto por el que se envía la BPDU. Los bridges que las reciben generan con estos datos más los costes de sus enlaces una BPDU propia que se envía a su vez a los bridges vecinos. Este algoritmo no dispone de un mecanismo de detección de terminación o convergencia del mismo en la red sino que utiliza temporizadores que esperan el tiempo suficiente para asegurar que haya convergido. STP por ello está basado en temporizadores fijos de 15 segundos para cambiar de estado los puertos, por lo que adolece de lentitud en la reconfiguración del árbol. La fiabilidad y disponibilidad se reducen con el número de nodos ya que la probabilidad de reconfiguración aumenta al aumentar el número de nodos. En la Figura 1 se muestra el diagrama de estados de puerto en STP y las transiciones permitidas. Tras la inicialización los puertos pasan a estado de bloqueo. Los estados estables en operación son Blocking y Forwarding. Los estados transitorios que atraviesa un puerto desde el estado Blocking hasta Forwarding son Listening and Learning. Cuando el algoritmo de

árbol de expansión deduce que un puerto debe pasar a estado forwarding primeramente lo pasa a estado Listening en que el puerto continua bloqueando el reenvío de tramas, pero recibe las BPDUs y las pasa al dispositivo para su proceso, pero no aprende direcciones MAC ni transmite BPDUs. Tras vencer el temporizador Forward Delay el puerto pasa a estado learning en el que el puerto aprende direcciones MAC, pero sigue sin reenviar tramas de usuario. El vencimiento del temporizador Forward Delay causa el cambio de estado a Forwarding, en el puerto además de aprender direcciones MAC y recibir y transmitir BPDUs reenvía tramas de usuario. La caducidad de la BPDU recibida está controlada por el temporizador Max Age, período de validez de cada BPDU recibida.

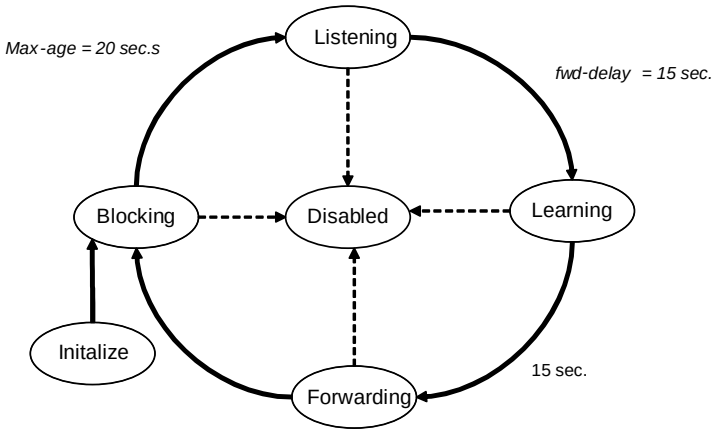


Figura 1. Estados de puerto en STP

2.1.2 Protocolo de Árbol Rápido de Expansión (RSTP)

El Protocolo Rápido de Árbol de Expansión (RSTP) fue estandarizado en 2001, inicialmente como 802.1w, enmienda al estándar IEEE 802.1D de 1998 y en 2004 ha sido incorporado al estándar 802.1D reemplazando a STP. RSTP está implementado alrededor del vector de prioridad de puerto, vector que incluye la identidad de bridge raíz, identidad de bridge designado, coste al bridge raíz, prioridad de puerto e identidad de puerto, vector que expresa su prioridad para ser elegido en el árbol de expansión. RSTP construye un árbol de expansión mediante anuncios de tipo vector distancia que contienen el coste al bridge raíz del árbol. En cada LAN, las BPDUs las transmite el puerto Designado. El puerto adquiere el rol de Designado al inicializarse o tras la elección de estado de puerto, para ello el puerto en cada extremo de la LAN compara su vector de prioridad con el recibido por la LAN por si debe pasar a ser designado. Se transmiten BPDUs siempre que cambia la información a transmitir, con una velocidad máxima configurable, además de periódicamente (Hello Time). Cada bridge compara a nivel de puerto todos los vectores de prioridad recibidos de los vecinos y elige el mejor vector al bridge raíz. Se define como BPDU *superior* si el identificador (bridge ID) de bridge raíz es menor o el coste del camino al raíz es menor. RSTP reemplaza el mecanismo basado en temporizadores de 15 segundos empleado por STP para verificar que el algoritmo ha convergido con un mecanismo de propuesta-aprobación controlado localmente entre conmutadores vecinos para el cambio al estado “forwarding” de los puertos en la dirección descendente del árbol de distribución. Este mecanismo requiere el uso de enlaces dedicados para funcionar correctamente sin crear bucles y se describe en

la Figura 2. Mediante este mecanismo, el enlace entre el puerto raíz y su bridge designado se habilita (el puerto pasa a forwarding) si previamente el bridge ha bloqueado todos los puertos designados, con lo que se va ampliando la topología activa nivel a nivel, a la vez que un "corte" horizontal en la red se propaga desde el bridge raíz hacia abajo al bloquearse los puertos designados.

Otra característica propia de RSTP es que todos los conmutadores emiten autónomamente las BPDUs cada Hello Time, en lugar de hacerlo tras la recepción de una BPDU del bridge raíz. Varios mecanismos adicionales contribuyen a obtener una reconfiguración rápida: cuando un conmutador recibe una BPDU con información menos preferente (es decir mayor identidad de switch raíz (Bridge ID) o coste de camino (métrica) al mismo), el switch receptor responde de inmediato al vecino con su propia BPDU para propagar la información superior de la que dispone. Con este mecanismo, un bridge que pierde su conexión con el bridge raíz recibirá rápidamente información de BPDU de los bridges vecinos que aún conservan su conexión al bridge raíz, seleccionando el que ofrezca la BPDU superior a efectos de elegir el nuevo puerto raíz. Aunque RSTP requiere enlaces punto a punto para el mecanismo de paso a forwarding, no los requiere para la transición rápida de puerto raíz, entre el puerto que era raíz y el preseleccionado de reserva (rol de Alternate).

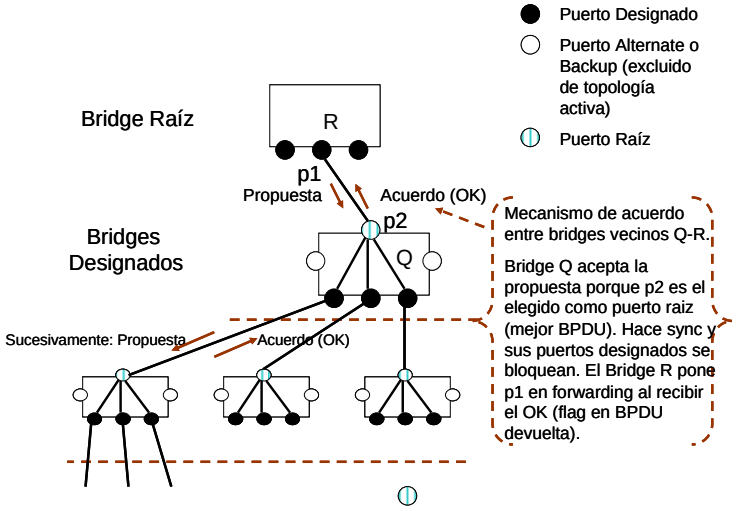


Figura 2. Mecanismo de RSTP para cambio de estado de puertos a Designado.

Los estados de puerto, que eran cinco en STP, en RSTP quedan reducidos a tres: *discarding*, *learning* y *forwarding.* El estado *discarding* agrupa a los estados *disabled, blocking* y *listening* de STP. En STP existía mezcla entre los conceptos de rol y el estado actual de puerto. En RSTP el rol de cada puerto es una variable del puerto. El rol de puerto está vinculado al estado del puerto. Existen dos roles nuevos de puerto ligados al estado *blocking* (en sentido STP) del puerto: *alternate port* y *backup port.* Estos estados de bloqueo del puerto se mantienen porque el puerto recibe mejores BPDUs de forma continuada por la LAN a la que está conectado, procedentes del mismo bridge (rol de *backup port*) o de distinto bridge *(*rol de *alternate port).* En caso de fallo del puerto raíz la elección del puerto alternate permite una conectividad alternativa al bridge raíz inmediatamente disponible por estar precualificado para ser puerto raíz.

Este criterio de BPDU superior recibida es el elemento base de decisión en RSTP. Se comparan las BPDUs recibidas por diferentes puertos o con la información

almacenada en el puerto. El puerto que recibe la BPDU superior asume el rol de puerto raíz (*root port).* Los puertos que pueden difundir la BPDU superior en el segmento de LAN al que están conectados asumen el rol de puertos designados.

RSTP envía BPDUs cada Hello Time y las utiliza para hacer caducar la información recibida rápidamente (aparte de utilizar el temporizador Max Age) y para detectar caídas en la comunicación entre bridges vecinos. Si un bridge no recibe ninguna BPDU en tres intervalos Hello Time considera caída la comunicación con su vecino. La detección es más rápida y precisa que en STP dado que se detecta exactamente el enlace en fallo.

La Figura 3 muestra el formato estándar de BPDU para RSTP y el detalle del octeto de indicadores (flags). Estos indicadores se implementan con un bit de información cada uno y los bridges vecinos los emplean para comunicar y asentir los roles asignados a los puertos por el protocolo, estados de puerto y las transiciones de estado de los puertos.

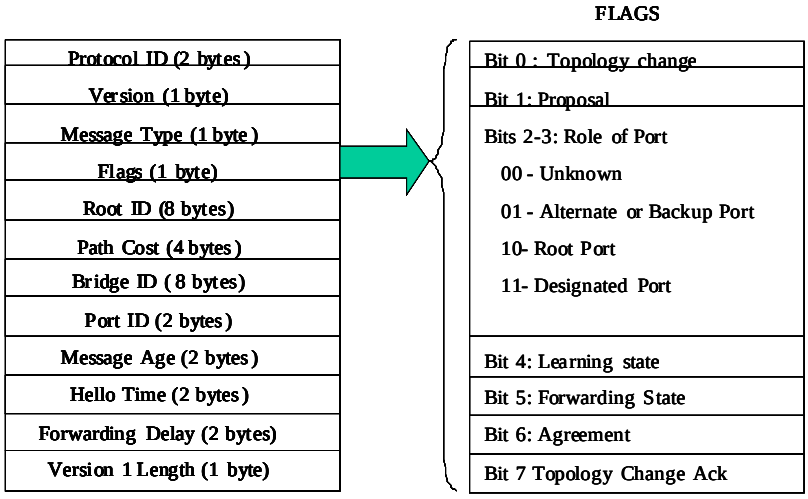


Figura 3. Formato de BPDU RSTP y detalle del octeto de indicadores (IEEE 802.1D).

Es destacable también un mecanismo proactivo para acelerar la reconfiguración consistente en que un puerto conectado a un puerto raíz del bridge contiguo, si recibe por el mismo una BPDU inferior que la suya, le contesta directamente con su información.

Los cambios de topología se comunican en la misma BPDU mediante el octeto de indicadores (flags) en vez de usar las BPDUs de cambio de topología (Topology Change Notification (TCN)). Al enterarse de un cambio de topología el bridge lo propaga por todos los puertos excepto aquel por el que se ha recibido y se procede al borrado inmediato de las direcciones MAC aprendidas (por ese puerto).

El mecanismo de propagación de los cambios de topología es asimismo mas rápido y completo que en STP, se propaga hacia abajo (además de continuar hacia arriba por el árbol de expansión) de inmediato desde cada nodo alcanzado al propagarse hacia arriba, en vez de propagarse primeramente hacia arriba hasta el bridge raíz y después hacia abajo por todo el árbol. RSTP es compatible con el protocolo de Agregación de Enlaces (enlaces paralelos en reparto de carga) ya incorporado en el estándar IEEE 802.3. Esto

permite el reparto de carga entre varios enlaces paralelos sin que sean cortados por el protocolo y posibilitando el fallo de uno de ellos sin impacto adicional a la reducción de la capacidad total del grupo de enlaces agregados.

En la Figura 4 se muestra un esquema simplificado de las máquinas de estado empleadas por el protocolo RSTP [802.1D]. Los bloques de la figura corresponden a máquinas de estado interrelacionadas. En la esquina derecha puede verse la máquina de estados de migración de protocolo por puerto que controla la ejecución de STP o RSTP en los puertos. Cada máquina de estados se implementa por cada puerto, salvo las de Selección y Transición de Roles de Puerto, que es única por bridge.

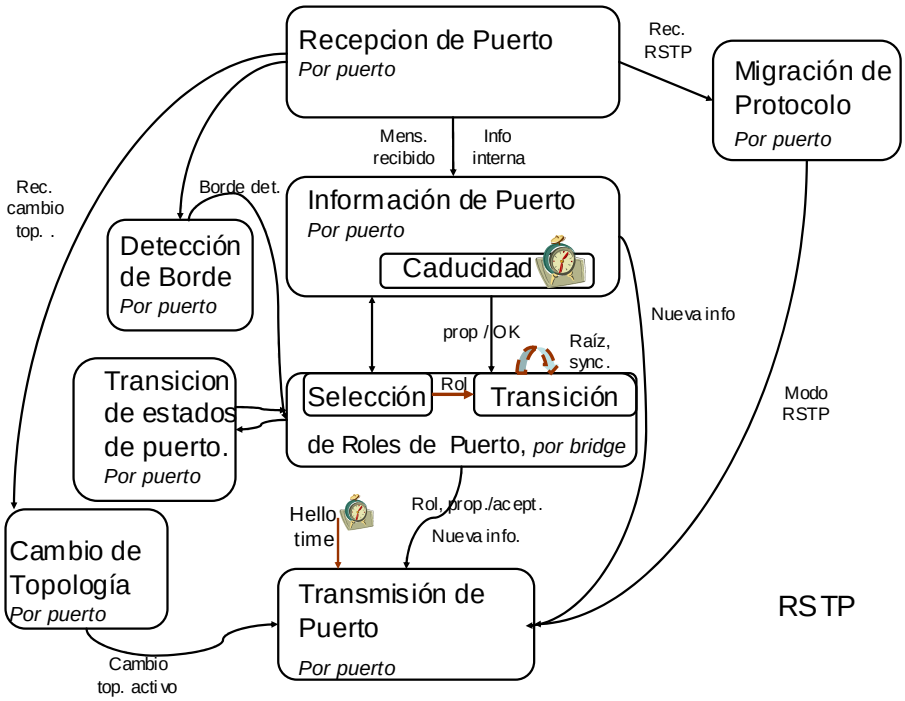


Figura 4. Relación entre máquinas de estados de RSTP.

RSTP presenta tiempos de reconfiguración muy pequeños, si existe un puerto preseleccionado para suceder al puerto raíz solamente requiere el tiempo de detección de caída del enlace (tan corto como 10 mseg). Si no existe puerto preseleccionado, se requiere solamente la transmisión de una BPDU en cada sentido entre los puertos del nuevo enlace que va a ser raíz. No obstante, RSTP puede presentar un comportamiento de cuenta a infinito en caso de caída del bridge raíz, resultando en ese caso retardos de hasta varios segundos [MEZ04].

2.1.2.1 Métricas aplicadas

El estándar IEEE 802.1D especifica los costes por defecto a aplicar a los enlaces en la construcción del árbol de expansión. Se muestran en la Tabla 1. Estos costes *c* van en relación inversa a la velocidad *v* del enlace según la ecuación $c * v = 20 * 10^{12}$. Utilizando las velocidades Ethernet de 10, 100, 1000, 10.000 Mbps en una red campus se tienen variaciones de cuatro órdenes de magnitud dentro de la misma red.

Como consideración aparte, estas diferencias de velocidad en los diferentes tramos se reflejarán en la desigual contribución de cada tramo capa en el retardo total de la red, el nivel de distribución y acceso será normalmente el causante de la mayor parte del retardo.

Tabla 1. Costes de enlace por defecto (IEEE 802.1D).

Link Speed	Recommended value	Recommended range	Range
<=100 Kb/s	200 000 000*	20 000 000–200 000 000	1–200 000 000
1 Mb/s	20 000 000[a]	2 000 000–200 000 000	1–200 000 000
10 Mb/s	2 000 000[a]	200 000–20 000 000	1–200 000 000
100 Mb/s	200 000[a]	20 000–2 000 000	1–200 000 000
1 Gb/s	20 000	2 000–200 000	1–200 000 000
10 Gb/s	2 000	200–20 000	1–200 000 000
100 Gb/s	200	20–2 000	1–200 000 000
1 Tb/s	20	2–200	1–200 000 000
10 Tb/s	2	1–20	1–200 000 000

2.1.3 Configuración de Bridges en IEEE 802.1D

Se pasa revista a continuación a los aspectos de gestión de bridges relativos exclusivamente a la configuración de los mismos. Se tratan estos aspectos porque uno de los objetivos de diseño es obtener la mayor autoconfiguración posible en los bridges de la red campus.

Las funciones de gestión de los bridges están definidas en la sección 14 del estándar IEEE 802.1D [802.1D04]. El área de Configuración es una de las áreas funcionales de Gestión junto con otras como Seguridad, Fallos y Prestaciones. La Gestión de Configuración posibilita la identificación de recursos de comunicaciones, inicialización y cierre de los mismos, aplicación de parámetros de operación, el descubrimiento de recursos y el establecimiento de relaciones entre los mismos. La gestión se especifica mediante entidades de gestión, entre las que destacan: la entidad de gestión de Bridge, las entidades individuales MAC de cada puerto del Bridge, el Proceso de Reenvío y la Base de Datos de Filtrado de la entidad de MAC Relay, la entidad del Protocolo de Árbol de expansión, y las de los participantes de los protocolos GARP y GMRP.

2.2 Protocolos de Árbol de Expansión con Caminos Alternativos.

Se agrupan aquí los antecedentes de protocolos propuestos que extienden o complementan el protocolo de árbol único de expansión mediante la utilización de enlaces adicionales al árbol de expansión, mejorando así la utilización y las prestaciones de la red

2.2.1 DLS/GDLS

La propuesta conocida como Distributed Load Sharing (DLS) [HAR89] es un antecedente de STAR [Lui02]. DLS permite utilizar algunos enlaces deshabilitados por el árbol de expansión. Para ello los enlaces deben cumplir varios requisitos: los bridges de ambos extremos del enlace deben ser DLS, no deben tener relación de bridge antecesor uno del otro en el Árbol de Expansión, y la longitud del enlace debe ser menor que la suma de las distancias al raíz de los dos bridges DLS. Una carencia de DLS es que puede sobreestimar el camino a través del árbol entre dos bridges DLS por lo que puede elegirse un camino alternativo entre bridges DLS más largo que el camino a través del árbol.

La propuesta de DLS fue modificada por Perlman [PHL92] y bautizada como DLS Generalizado (GDLS), haciéndola más simple y permitiendo usar cualquier enlace que no forme parte del árbol de expansión que termine en bridges GDLS. En lugar de comparar la longitud del enlace cruzado (extra-árbol) con el camino correspondiente por el árbol, se compara la velocidad de ambos caminos, la cual se obtiene enviando uno de los bridges al otro una BPDU especial por el camino del árbol y otra por el camino extraárbol. Esto requiere mantener esa información por cada enlace aunque finalmente no se utilice dicho enlace como camino adicional al árbol de expansión. Con GDLS no se puede garantizar que el camino extraárbol no es peor que el camino de árbol para métricas aditivas, salvo si dicha métrica es el retardo. GDLS es compatible con bridges 802.1D.

2.2.2 STAR

El protocolo de Árbol de Expansión Alternativo (Spanning Tree Alternate Routing Protocol (STAR)) [Lui02] se concentra en redes campus de tamaño moderado con un único prefijo de subred IP y mínima configuración. STAR permite utilizar caminos alternativos (a través de enlaces cruzados o *cross links*) a los del árbol de expansión entre bridges STAR. Incluye mecanismos para descubrimiento mutuo de los bridges que implementan el protocolo y posibilita el uso de caminos más cortos que los del AE cuando existen. No está orientado a redes troncales sino a LANs y es compatible con los bridges estándar 802.1D. Los caminos pueden ser mejores que los de STP pero no son necesariamente mínimos. Dada su compatibilidad con STP, es posible, mediante algunos Bridges STAR insertados en una red con bridges estándar que ejecutan STP, habilitar enlaces transversales entre Bridges STAR, enlaces que con STP quedarían bloqueados para evitar los bucles. Se modifica el protocolo STP introduciendo un estado adicional

STAR a la secuencia normal de iniciación y actualización de STP. Se utiliza encapsulado adicional de nivel 2 entre Bridges STAR para lograr reenviar las tramas STAR a través de los Bridges normales STP intermedios, dado que de no encapsular podrían ser descartadas las tramas con un campo Type desconocido por los Bridges 802.1D. El encaminamiento se basa en Vectores Distancia.

Adicionalmente a STAR se han propuesto en [Lui02] protocolos adicionales para realizar un posicionamiento óptimo de los bridges STAR en la red (protocolo PEARL) y extensiones de STAR para elección de caminos alternativos al árbol de expansión según requisitos de calidad de servicio (QoS) [LLN03] que extiende el protocolo STAR para caminos múltiples posibilitando la elección del camino estándar del árbol de expansión, el más corto (atajo o enlace cruzado, *cross link)*, o caminos híbridos, (parte vía AE y parte con enlaces cruzados). Se asignan unos u otros según el valor del campo de prioridad (tres bits) en el octeto adicional 802.1Q (prioridad de 0 a 7) de la trama y según el número de saltos ya efectuados por la trama (lo que eleva su prioridad relativa). Se estudia el rendimiento para redes de 20 a 30 nodos y la mejora de longitud de camino para primera prioridad es modesta (1,3) si bien se afirma que la mejora aumentará en redes grandes al haber mayor número de caminos alternativos en promedio. Se considera aplicable asimismo para reparto dinámico de carga en las redes.

2.2.3 RSTAR

El protocolo RSTAR que se describe en esta sección es la lógica migración de STAR a RSTP que propone el autor de esta Tesis. Es la actualización del protocolo STAR mencionado en la sección anterior para utilizar el recientemente estandarizado protocolo básico RSTP como base en lugar de STP, el protocolo anteriormente estándar en 802.1D, cuando STAR fue propuesto. El protocolo STAR incluye STP, pudiéndose considerar como una extensión del mismo por lo que no se habla separadamente de RSTP sino del protocolo RSTAR como resultado de extender RSTP con STAR.

2.2.4 Procesos de RSTAR

Siguiendo el modelo de STAR descrito en [Lui02], RSTAR incluye los procesos del protocolo RSTP en lugar de los de STP más los específicos de STAR para encaminamiento por enlaces cruzados.
Los procesos principales que lo componen son:

- Elaboración del árbol RSTP de acuerdo al estándar IEEE 802.1D [IEEE1D04].
- Procesos estándar RSTP de reenvío de tramas y aprendizaje de direcciones MAC en puertos.
- Procesos básicos de Aprendizaje y Reenvío RSTAR.

2.2.5 Procesos básicos RSTAR

El protocolo RSTAR consta de los siguientes procesos:

- Proceso de Búsqueda de caminos RSTAR

Este proceso calcula la tabla de encaminamiento del bridge (*TEB*). Esta tabla, si es posible, contendrá el bridge RSTAR de siguiente salto y puerto de salida del camino más corto para cada bridge RSTAR destino. A su vez consta de tres procesos:

- Descubrimiento de Vecinos RSTAR Distantes. El objetivo de este proceso es descubrir otros bridges RSJ y obtener distancias exactas o aproximadas.
- Descubrimiento de Vecinos RSTAR Directos. Su objetivo es descubrir a los bridges RSJ vecinos inmediatos y obtener y actualizar distancias a ellos.
- Mejora de Vector Distancia. Objetivo: Obtener rutas a todos los bridges RSTAR.

Las descripciones de estos procesos para el protocolo STAR [Lui02] son aplicables a RSTAR.

- Proceso de Aprendizaje RSTAR. En el protocolo STAR original, este proceso consiste en el aprendizaje de los sistemas finales asociados a cada bridge (de los que el bridge es *designado*) mediante intercambio de listas de sistemas finales. En RSTAR, por utilizarse servidores ARP, no es necesario intercambiar estas listas de sistemas finales dado que el servidor ARP facilita el AB designado del sistema final destino.

- Proceso de Reenvío RSTAR (Forwarding). Este proceso consiste en el envío de tramas entre dos bridges RSTAR usando caminos alternativos encontrados mediante el protocolo RSTAR. Utiliza la tabla de reenvío (forwarding) calculada por el proceso *Búsqueda de caminos RSTAR* . No presenta modificación respecto a STAR.

- Resolución ARP, AB y registro. El proceso de resolución ARP, resolución de AB y registro para ARP se describe en detalle en el epígrafe del escenario de encaminamiento.
- Resolución de sistema final y ABs (aprendizaje de sistemas finales en STAR). Dado que se utilizan servidores ARP en los ABs, no hay tal proceso de aprendizaje de sistemas finales. Este proceso utiliza un Bridge Agente (AB). Bridge Agente se define como el bridge RSTAR antecesor en el árbol de expansión que es más cercano al bridge correspondiente. Dado que una trama BPDU que emita un bridge RSJ será escuchada por sus vecinos, cuando un bridge RSTAR recibe una BPDU de un vecino sabe si es RSTAR porque emite BPDUs con identidad de protocolo RSTAR a la dirección multicast de bridges RSTAR. Este Bridge Agente es el responsable de la difusión y manejo de todos los sistemas finales situados debajo de él en el árbol de expansión.

RSTAR utiliza las BPDUs de cambio de topología (TCN BPDUs) difundidas por la red para proceder al borrado de las tablas y reaprendizaje.

2.3 Protocolos de Árboles Múltiples de Expansión.

Se describen en esta sección los protocolos que utilizan más de un árbol de expansión para difundir las tramas en la red conmutada.

2.3.1 Árboles múltiples. Multitree.

Sincoskie y Cotton [SC88] propusieron *Multitree* (MT), un algoritmo de árboles múltiples que destaca porque permite utilizar toda la red instalada, creando más de un árbol de expansión y asignando cada bridge al menos a un árbol de expansión.

Un aspecto básico en los protocolos de árboles múltiples de expansión es el criterio de asignación de paquetes a árboles. En el caso de Multitree, la asignación de un paquete a un árbol se realiza mediante una función que combina las direcciones MAC origen y destino en un número único de árbol (p.e. función OR exclusivo de ambas direcciones MAC). Si todos los bridges utilizan la misma función, el paquete será encaminado en todos los bridges por el mismo árbol. El número de árbol debe guardarse junto con la identidad de nodo para poder hacer reenvío de paquetes. La asignación de paquetes a árboles puede ser explícita o implícita. En el primer caso se incluye el identificador del árbol en el paquete y en el segundo se aplica una función.

Multitree propone interesantes mecanismos activos en los sistemas finales para el mantenimiento de los árboles multicast. Los sistemas finales mantienen los árboles multicast enviando periódicamente paquetes de refresco para mantener las direcciones activas en las cachés. Este mecanismo ha sido utilizado posteriormente en gestión de movilidad en capa dos.

Aunque la utilización de la red con Multitree es muy alta, la longitud de camino relativa puede ser grande porque la función que asigna paquetes a árboles es aleatoria y puede asignar los paquetes a árboles de poca efectividad. Tanto la obtención de una función que optimice las longitudes de camino asignando los bridges a los árboles más adecuados como el problema de repartir la carga entre los árboles quedan sin resolver.

Es de interés mencionar la propuesta [RF91] en la que se emplean *Source Dependent Spanning Trees (SDS),* árboles de expansión creados dinámicamente en función del sistema final. Se utiliza un árbol de expansión por defecto, pero mediante comparación de retardos de los paquetes del sistema final, se construye un árbol SDS que lo reduce. Su complejidad es alta por manejar para el encaminamiento parejas de sistemas finales en vez de costes.

2.3.2 Árbol de Expansión por VLAN (PVST).

Tras la definición de las redes locales virtuales (VLAN) y su estandarización en el estándar IEEE 802.1Q [VLAN] apareció la posibilidad de que estas nuevas redes

virtuales montadas sobre una única red física, pudieran estar soportadas por diversos árboles de distribución, adecuados a su topología y aplicación. Para ello se desarrollaron diversos protocolos propietarios tales como Per VLAN Spanning Tree (PVST y PVST+) [Cisco] que permiten configurar diferentes árboles de distribución y asignarlos a diversas VLAN. Esta asignación se realiza por VLAN, por lo que puede consumir excesivos recursos para mantener en el conmutador la asignación de hasta 4094 instancias posibles de árbol, correspondientes a la asignación de un árbol por VLAN. La posterior estandarización del protocolo de árboles múltiples MSTP permitió que varias VLAN compartan un único árbol de expansión con el consiguiente ahorro de recursos.

2.3.3 Protocolo de Árbol de Expansión Múltiple (MSTP)

Se describe a continuación el Protocolo de Árbol de Expansión Múltiple (Múltiple Spanning Tree Protocol (MSTP) [802.1D]. Este protocolo está basado en la configuración de instancias múltiples simultáneas de árboles de expansión en una determinada región. El protocolo MSTP provee de conectividad simple y completa a las tramas asignada a cualquier VLAN de una red local conmutada compuesta por bridges MSTP o RSTP. MSTP permite utilizar caminos separados a las tramas pertenecientes a diferentes VLAN, utilizando instancias independientes MST, instancias válidas dentro de cada región MST formada por LANs y bridges MST.

MSTP ha sido estandarizado en 2003 dentro del estándar para VLANs en bridges IEEE 802.1Q. Anteriormente existió como IEEE 802.1s (draft) Multiple Spanning Tree, como enmienda al estándar 802.1Q. MSTP es la extensión de RSTP para árboles de expansión múltiples. Esto permite una mejor utilización de la infraestructura y distribución de la carga, así como mejor resistencia a fallos por la multiplicidad de árboles construidos, al precio de la complejidad de configuración. Prevé diversos dominios de administración con lo que las áreas MST pueden ser gestionadas por diferentes administradores. Existe un árbol común CST y árboles por regiones, por lo que en caso de reconfiguración solo se afecta esa región y los tiempos de convergencia se reducen en redes grandes.

MSTP se apoya parcialmente en el protocolo RSTP, lo requiere y es compatible con él. Asimismo utiliza los mecanismos definidos para VLAN. Construye un conjunto de árboles de distribución independientes (instanciaciones o instancias) (MSTI) en una región MST de la red. Dicha región se conecta a otras regiones MST, a efectos del control de protocolos STP, mediante un Árbol de Expansión Común (Common Spanning Tree (CST)) como se muestra en la Figura 5.

Dentro de una región varias VLAN pueden asignarse a una sola instancia de árbol. Mediante el uso de múltiples instancias de árbol en la misma región, es posible un mejor aprovechamiento de los enlaces, utilizando enlaces que son bloqueados por el algoritmo de spanning tree. Dentro de cada región se mantiene su continuidad interna mediante un Árbol de Distribución Interno (IST), identificado como instancia 0 (IST 0), que actúa como árbol básico de distribución dentro de la región. El CIST o árbol total se compone del CST que conecta las regiones y del IST que realiza la conectividad dentro de cada región. Ello permite la gestión independiente de las regiones.

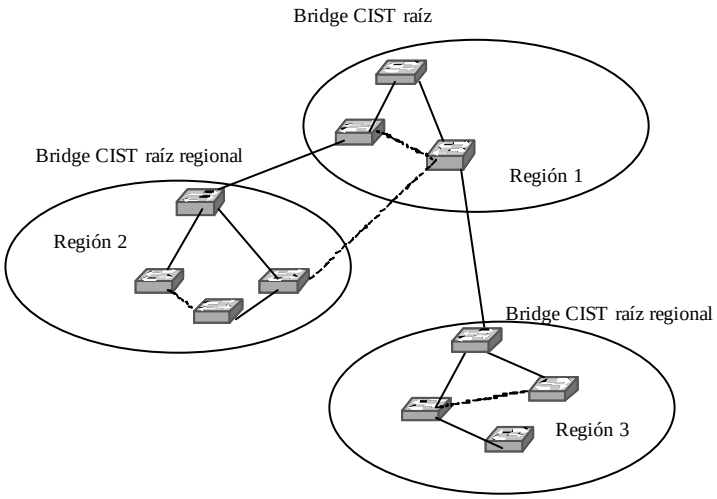


Figura 5. Regiones MST y Bridges Raíz Regionales en el árbol total CIST

La región aparece hacia el exterior como un único *superbridge* independiente. Es decir, toda la región se conecta al CST a través de un Puerto Raíz Regional y un cierto número de Puertos Designados, como si estuviera formada por un único bridge. Mediante mecanismos específicos se evita que el árbol IST salga de la región y vuelva a entrar en la misma. De esta forma los cambios externos de topología no provocan cambios internos. El bridge raíz del CST es la raíz de toda la red. Por cada región MST hay un árbol IST maestro, el cual pertenece al CST. MSTP es adecuado para capa de distribución de redes campus, algo menos para troncales debido a la complejidad de configuración de las VLANs, aunque existen algunas propuestas recientes de sistemas de delimitación y configuración automática de regiones.

2.3.4 Configuración de bridges en IEEE 802.1Q

En cuanto a la especificación IEEE 802.1Q en lo relativo a gestión de topología de VLANs, contempla lo siguiente:

- Configuración estática de VLANs mediante mecanismos de gestión.
- Configuración dinámica de VLANs mediante generación y distribución de información de pertenencia con el protocolo GVRP.

La configuración y gestión de VLANs supone una ampliación de las funciones de filtrado especificadas para RSTP en cuanto al control de configuración de las instancias de árboles MSTI, reinicialización de los mismos, prioridades, pertenencia a VLANs, etc.

Uno de los aspectos clave de la configuración de MSTP consiste en establecer la correspondencia entre VLANs e instancias de árboles MSTI. Como consecuencia de ello, la configuración de MSTP es compleja y delicada. Debe configurarse en cada bridge la correspondencia entre VLANs (hasta 4096) e instancias de árbol de distribución (hasta 64

en el estándar). MSTP utiliza múltiples identidades de bridge (Bridge ID), dado que requiere una por cada instancia de árbol. Para reducir el número de direcciones MAC necesarias, se puede emplear el mecanismo denominado de reducción de direcciones MAC, en el que el campo de prioridad del bridge, (formado por los dos octetos más significativos de la Bridge ID), se descompone en dos partes: los 4 bits más significativos y los restantes 12 bits. Los 4 bits se usan para codificar la prioridad del bridge entre 0 y 16, los 12 bits transportan el número de VLAN (VLAN ID). Los seis bytes restantes de la Bridge ID contienen, como de costumbre, la dirección MAC del bridge, con lo que se obtiene una bridge ID única para cada instancia de árbol.

2.3.5 Global Open Ethernet

Global Open Ethernet (GOE) [Iwa+04] es un proyecto de NEC para redes privadas virtuales de capa dos, reivindicado como autoconfigurable, que se propone como alternativa al encapsulado anidado denominado Q-in-Q (anidamiento sucesivo de etiquetas 802.1Q de VLAN) y a Ethernet sobre MPLS (EoMPLS) por ser ambos sistemas complejos de gestionar. Q-in-Q es además poco eficiente en el uso de la red y los equipos EoMPLS son muy costosos. GOE utiliza conmutación de etiquetas (*tag switching*) para encaminar por la red del proveedor. Se asigna una VLAN ID a cada bridge, para de esta forma utilizar la etiqueta de VLAN asignada (VLAN tag) como dirección de encaminamiento (lo mismo que hace MSTP con las tramas por los árboles de expansión), lo cual en principio es mas parecido a IP que lo que hace MPLS.

En GOE se utiliza, como se muestra en la Figura 6, una etiqueta jerárquica de longitud variable (compatible con la de VLAN). Esta etiqueta GOE tiene una parte obligatoria y otras dos opcionales (Customer ID y Vendor ID), desacoplando los clientes y la respectiva VLAN de usuario. La parte obligatoria de la etiqueta se utiliza como dirección de nodo destino para el encaminamiento en los árboles, con la misma longitud que una identidad de VLAN. Como la etiqueta de VLAN estándar es de 12 bit, se produce una compresión del espacio de direcciones de 48(MAC) +12 (VLAN) bits a 12 bits. Esto simplifica y acelera el procesado de tramas en los bridges. GOE es compatible con el protocolo MSTP estándar. Dado que la dirección del switch destino se codifica en 12 bit en la etiqueta obligatoria VLAN, en ello se basa la compatibilidad con MSTP. La jerarquía es recursiva y se permite apilamiento sucesivo de las VLAN y de los árboles RSTP/MSTP. Debido a la compatibilidad con MSTP, es presumible que la configuración de los bridges sea compleja, pero desconocemos la implementación. Se describe GOE porque, aunque las L2VPNs tienen un escenario y requisitos diferentes al planteado, se considera que pueden aplicarse parcialmente algunos conceptos de las redes metropolitanas y L2VPNs, tales como el etiquetado adicional para el encapsulado para reducir el número de direcciones MAC y el empleo de árboles múltiples de expansión para el encaminamiento. Por otra parte, la compatibilidad entre redes campus, MAN y L2VPNs requiere que las soluciones puedan ínter operar.

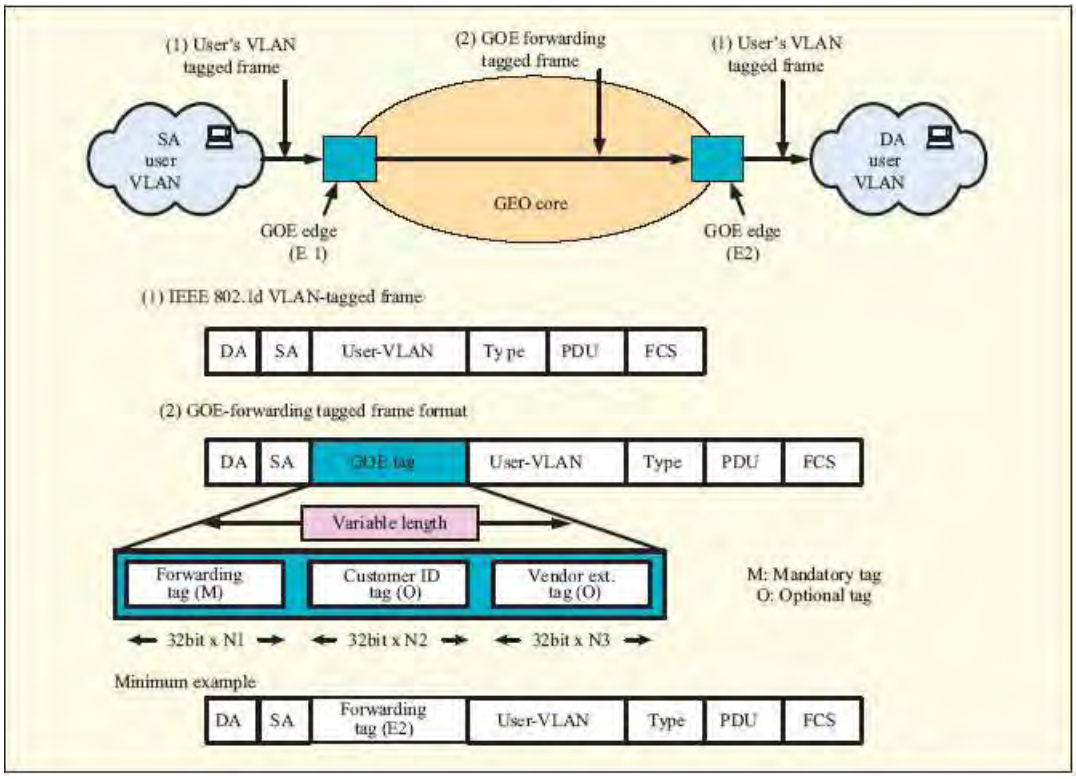


Figura 6. Etiquetado jerárquico de tramas en Global Open Ethernet (GOE)

GOE se inscribe dentro de las estrategias para suprimir funcionalidad de encaminamiento e ingeniería de tráfico de la capa IP trasladándola a la capa 2 (como MPLS y GMPLS).

2.3.6 VLANs automáticas calculadas (Viking).

La complejidad de configuración que por un lado presenta MSTP y las posibilidades para aprovechamiento de la infraestructura de la red que por otro brindan los árboles múltiples de expansión están haciendo surgir propuestas que utilizan configuración externa de las VLAN y de los árboles MSTP en los bridges mediante SNMP. Se analiza a continuación la denominada Viking [Sha+04].

Viking es una arquitectura orientada a redes metropolitanas y redes de tipo *cluster* como las redes de almacenamiento *(Storage Area Networks SAN).* Utiliza árboles múltiples de expansión para lograr la utilización eficiente de la infraestructura. Los dispositivos que intervienen son el *Viking Manager*, elemento central de la red, los bridges y los sistemas finales. El *Viking Manager* precalcula las rutas entre sistemas finales, tanto activas como las rutas alternativas de reserva. Para optimizar la utilización de la red, las rutas se agregan asignándolas a una VLAN determinada. La configuración de las VLANs y el árbol de expansión correspondiente en los bridges de la red la realiza el *Viking Manager* mediante comandos SNMP. Existe un árbol de expansión por cada VLAN, el cual se forma agregando caminos de rutas calculadas entre nodos. Viking requiere que el comportamiento de los sistemas finales actuales se modifique, dado que los sistemas finales incorporan una lógica de selección de VLAN en la pila de protocolos de red. Los conmutadores Ethernet son estándar, aunque se estudian modificaciones en los mismos para optimizar el funcionamiento. La detección de fallos se realiza mediante

traps SNMP. El diseño del sistema impide una respuesta rápida para reconfiguraciones debido al tiempo necesario para la detección del fallo (400-600 mseg.) estando en menos de un segundo el tiempo total de indisponibilidad de la red, que aunque es mejor que el caso peor de RSTP, no supone una mejora radical sobre el mismo. La utilización de interfaces SNMP para configuración automatizada de MSTP presenta las ventajas de la autoconfiguración (vista globalmente) y la compatibilidad con bridges estándar (que soporten MSTP). La necesidad de modificación de los sistemas finales hace inviable su aplicación en el contexto de esta Tesis.

2.3.7 IEEE Shortest Path Bridging

El autor no ha detectado actividad en IEEE durante los últimos años relacionada con el tema de esta Tesis, fuera de las actividades de estandarización normales de RSTP y MSTP (aunque no se ha realizado un seguimiento sistemático). Sin embargo, en marzo de 2005, bajo la denominación *Shortest Path Bridging* se ha presentado una propuesta, y se ha elaborado una solicitud de aprobación de proyecto [PAR05] para su votación en la reunión plenaria del Grupo 802 de Julio de 2005. El planteamiento básico de Shortest Path Bridging se describe a continuación:

- Se entiende por *shortest path bridging region* una región (en el sentido que tiene en el protocolo MSTP) en la que se establecen caminos mínimos entre los bridges (*shortest path bridges)* que pertenecen a ella mediante el establecimiento de árboles de expansión múltiples unidireccionales, con raíz en cada bridge del área. Para que el aprendizaje de las direcciones MAC funcione correctamente, es necesario :
  - Bien que el tráfico se difunda confinado por un único árbol, caso del árbol de expansión,
  - O bien que el tráfico de una estación *a* unida a *A* se confine por un árbol enraizado en A y el de una estación *b* se confine por un árbol enraizado en el bridge *B.* Es necesario que los árboles en direcciones opuestas entre dos bridges coincidan: el camino de *A* a *B* sea idéntico al camino de *B* a *A* por el árbol enraizado en *B.* De esta forma las direcciones aprendidas por un puerto pueden utilizarse como destino por el mismo puerto. Suponiendo una VLAN asociada a cada uno de dichos árboles, deben utilizarse las VLAN de forma bidireccional, y deben hacerse sinónimas en lo que se denomina *pairwise shared learning VLANs.*
- Se plantea una simplificación del protocolo estándar MSTP para implementarlo (sin definir).

En la propuesta de autorización [PAR05] se plantea el proyecto como una enmienda con la denominación Shortest Path Bridging al estándar 802.1Q. (VLANs). El objetivo del futuro estándar es, textualmente: “This standard specifies shortest path bridging of unicast and multicast frames, including protocols to calculate multiple active topologies that can share learnt station location information, and support of a VLAN by multiple, per topology, VLAN identifiers (VIDs)”. Se implementará mediante extensión de las especificaciones existentes de bridging para implementar bridging por caminos mínimos dentro de regiones definidas administrativamente, conservando las funcionalidades actuales. El objetivo es mejorar la utilización de la infraestructura y reducir la latencia. Se considera factible el actualizar mediante software muchos bridges de VLAN actuales. Existe fuerte apoyo de los fabricantes y gran interés de los usuarios en alternativas al

Árbol de expansión único. Los costes de los bridges se considera que no se verán afectados por la nueva funcionalidad y pueden ahorrarse encaminadores IP en muchos casos, reduciendo los costes de las redes. Desde el punto de vista de compatibilidad, debe mantenerse la compatibilidad de la enmienda a la 802.1Q que describimos aquí con los estándares IEEE 802.1D y anterior versión de 802.1Q. Como requisito destacable, los sistemas finales no deben requerir actualización alguna software o hardware. Para terminar, el proyecto se considera factible tanto técnicamente, basados en la experiencia de tecnologías VLAN, como económicamente.

El planteamiento inicial del problema es cercano en algunos aspectos a la presente Tesis según muestra la contribución inicial [SEA05] y lo que proponíamos en [IGA04], en el sentido de utilizar árboles múltiples, árboles cuyo bridge raíz está en el borde del área (*shortest path bridging region* en la terminología IEEE), en la necesidad para ello de utilizar protocolos simplificados respecto a MSTP, y en su aplicación a redes troncales de muy alta capacidad. Un problema pendiente es que si los árboles simétricos entre dos bridges tienen empate de coste pueden elegir caminos distintos y el aprendizaje de MACs no funciona correctamente. La propuesta existente al respecto para subsanarlo emplea los denominados *reflection vectors* [Fin05], que limitan el número de bridges participantes, alargan la trama y aumentan el tiempo de convergencia.

2.4 Protocolos de Prohibición de Giros.

Se describen en esta sección los protocolos basados en prohibición de giros, utilizan sistemas alternativos al árbol de expansión que buscan ser más eficientes en la prevención de bucles en la red de forma que se inhabiliten en menor grado los enlaces utilizados, prohibiendo determinados *giros* en los itinerarios por nodos de la red, en vez de inhabilitar enlaces.

2.4.1 Encaminamiento arriba/abajo *(Up/down routing*)

A continuación se expone el mecanismo de encaminamiento que utiliza Autonet, denominado encaminamiento arriba/abajo (*Up/Down routing)*. Este encaminamiento se basa en asignar un sentido a todos los enlaces de la red según la posición del vértice del enlace en el árbol de distribución (arriba si está mas cercano al Bridge Raíz, hacia abajo si al contrario, para lo cual se asignan identificadores crecientes a los bridges partiendo del bridge raíz y descendiendo nivel a nivel, como se muestra en la Figura 7.Los enlaces entre nodos a la misma altura reciben la orientación según la identidad del Bridge sea mayor o menor. El encaminamiento arriba/abajo evita los bucles porque un bucle siempre contiene dos giros abajo/arriba o arriba/abajo. Por lo tanto, si se obliga a que en cualquier recorrido, una vez que se ha girado hacia abajo, no pueda volver a girar hacia arriba, se evitarán los bucles en las tramas enviadas. Una ruta legal es la que nunca usa/atraviesa un enlace en la dirección hacia arriba después de haber usado uno hacia abajo. Es decir, prohibiendo los giros abajo-arriba se evitan los bucles. La efectividad del encaminamiento arriba/abajo depende principalmente de la elección del árbol de distribución, y en particular del Bridge Raíz. Para la topología típica de las redes campus estructuradas jerárquicamente esto tiene menor incidencia que para una red mallada

arbitraria. El problema del encaminamiento arriba/abajo es que no garantiza rutas óptimas, y que se hace menos eficiente a medida que la red se hace más compleja. Silla y Duato [SD97] utilizaron una variante de encaminamiento arriba/abajo en redes de estaciones de trabajo *(Networks of Workstations (NOW)).* Esta variante, de mejor rendimiento, se basa en la utilización de canales adicionales (virtuales) a los enlaces normales para aumentar las posibilidades de elección de rutas mínimas y evitar de esta manera bloqueos, pero subsisten los inconvenientes básicos del encaminamiento arriba/abajo convencional. El encaminamiento arriba/abajo mejora la efectividad del Árbol de Expansión, pero la congestión en las cercanías del Bridge Raíz permanece.

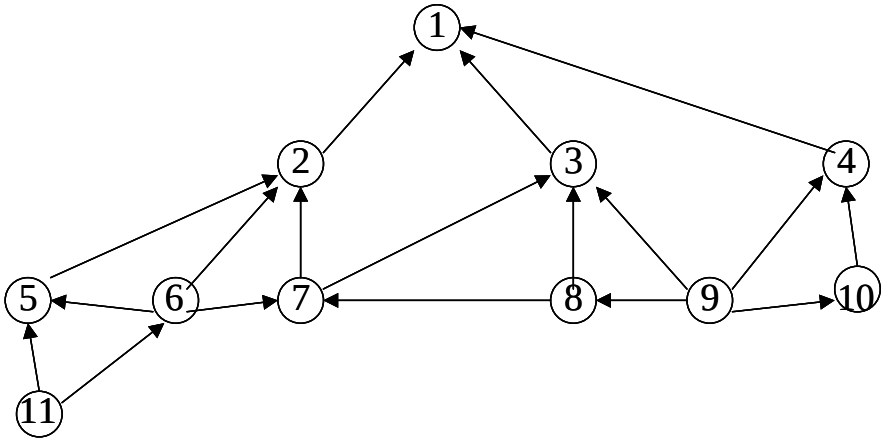


Figura 7. Orientación de enlaces en el encaminamiento Arriba/Abajo.

2.4.2 Algoritmos de Prohibición de Giros

Los algoritmos basados en Prohibición de Giros (*Turn Prohibition (TP))* constituyen una evolución del encaminamiento arriba/abajo. Una versión de interés de TP es la propuesta en 2002 en [SKZ02]. Dicho interés proviene de la oportunidad de poder aplicar el denominado Network Calculus a las redes, para lo que éstas deben ser de tipo *feed-forward*. Para garantizar esto, las redes deben estar libres de bucles y esto se logra desarrollando un algoritmo que prohíbe algunos *giros* en la red, en lugar de cortar enlaces completos, como hacen los protocolos de árbol de expansión como STP.

Este algoritmo opera de forma iterativa sobre la red nodo a nodo, seleccionando los giros a prohibir. Se han propuesto dos versiones, una básica (que puede resultar en subredes inconexas) y otra completa. El algoritmo suprime como máximo 1/3 del total de giros de la red. Para la versión ponderada del algoritmo, que valora la importancia de los enlaces mediante pesos, se demuestra que se prohíben menos de la mitad de los enlaces en media ponderada. Siendo el *grado* de un nodo de la red el número de sus enlaces, la complejidad es $O(N^2 d)$ siendo d el grado máximo de los nodos de la red. Los algoritmos de Prohibición de Giros operan normalmente en dos fases: en la primera se define el conjunto de giros prohibidos y posteriormente se construyen las tablas de encaminamiento. La definición de los giros prohibidos consta a su vez de tres fases: construcción del árbol de expansión, etiquetado de nodos según el árbol de expansión y algoritmo de definición del conjunto de giros prohibidos.

Otra propuesta posterior se denomina *Tree-Based Turn Prohibition Protocol (TBTP)* [Pel+04] y referencia los giros prohibidos o no con relación al árbol de expansión. La aplicación de la prohibición de giros en los algoritmos de encaminamiento como el de Djikstra (SPF) no es posible de forma simple (eliminando los caminos que contienen giros prohibidos de la tabla de encaminamiento) porque entonces los caminos no son mínimos debido a que Djikstra opera de forma escalonada. En [FE04] se estudia un reciente método para encaminar aplicando restricciones de prohibición de giros, pero la complejidad crece desde $O(N^2)$ a $O(E^2)$, siendo N y E respectivamente el número de nodos y de enlaces.

2.5 Bridges con encaminamiento

Se revisan en esta sección los diversos tipos de bridges que utilizan encaminamiento de diversos tipos. Se comienza por los bridges con encaminamiento en origen, se continúa con los bridges que ejecutan protocolos de encaminamiento centralizados y distribuídos y se termina con los bridges que incluyen en mayor o menor grado funcionalidad de encaminadores.

2.5.1 Bridges con encaminamiento en origen

Los bridges con encaminamiento en origen *(source routing bridges)* han representado desde su aparición una alternativa, basada en encaminamiento (aunque estandarizada como bridging), a los bridges transparentes [DP88] [PW87]. En el encaminamiento en origen, la ruta es establecida por el dispositivo origen e incluida en los paquetes enviados. Cada dispositivo intermedio observa su lugar en la ruta y encamina al siguiente dispositivo indicado en la ruta explícita. Los bridges con encaminamiento en origen más conocidos son los usados en redes de Paso de Testigo en Anillo (*Token Ring*), actualmente obsoletos. Uno de los inconvenientes de este tipo de bridges es, a diferencia de la autoconfiguración de los puentes Ethernet, la necesidad de asignar la identidad de todos los Puentes y Redes Locales. Dado que en el planteamiento del problema se ha establecido el requisito de no afectar a los sistemas finales, deberán ser siempre los bridges los que realicen el encaminamiento en origen, nunca los sistemas finales, para mantener la transparencia. El segundo gran inconveniente es la gran sobrecarga de mensajes producida por el descubrimiento de rutas. Aunque el encaminamiento en origen aparezca como obsoleto en redes fijas, es objeto de estandarización en las redes *ad-hoc* como es el caso del protocolo *Dynamic Source Routing (DSR)* [DSR04].

2.5.2 Autonet

El antecedente más destacado como propuesta de evolución de las tecnologías LAN es Autonet [SHO+91], ya que constituye uno de los primeros antecedentes de red local autoconfigurable. Autonet utiliza todo el ancho de banda disponible en sus enlaces empleando encaminamiento arriba/abajo (up/down), soportado sobre un protocolo de

árbol de distribución STP modificado. Las tramas Autonet se utilizan para encapsular las tramas Ethernet, y las direcciones empleadas son de 12 bits para simplificar el encaminamiento. La traducción de direcciones UID (direcciones MAC 48 bit) a direcciones cortas Autonet (12 bit) se realiza mediante aprendizaje de las tramas recibidas. Las direcciones cortas de los Host y Puentes son asignadas por el Bridge Raíz, que asimismo distribuye la información de topología y de árbol de distribución, de forma piramidal hacia abajo. Cada bridge calcula su tabla de encaminamiento con la información recibida. Un inconveniente considerable de Autonet radica en que la compatibilidad entre los modos de trabajo Ethernet y Autonet se implementa en los sistemas finales, requiriendo su modificación mediante la incorporación del módulo Localnet, situado por encima de los manejadores SW (drivers Ethernet y Autonet). Localnet se requiere para poder aprender la correspondencia entre las parejas de direcciones de formatos EUI/Autonet. Los enlaces deben ser punto a punto, con un solo sistema final conectado en cada segmento, de LAN, lo que facilita la localización de los sistemas finales, una parte del problema de encaminamiento.

2.5.3 Plano delgado de control / Plano de control distribuido.

En [MEZ04] y [Rex+04] se parte de la idea de que el modelo de servicio de Ethernet, basado en la difusión, no es adecuado para redes conmutadas y que la difusión realizada por protocolos como ARP y DHCP, reduce la fiabilidad. Se proponen dos alternativas: una basada en un plano delgado de control y la segunda en un plano de control distribuido. En cuanto a la primera, el plano de control se divide en *plano de decisión* y *plano de diseminación.* El plano de decisión calcula las rutas y el de diseminación es el encagado de enviar información de estado de la red al plano de decisión y entregar el resultado en forma de mensajes de configuración de los conmutadores. De esta forma el cálculo de las tablas de reenvío se traslada a un conjunto de servidores. El plano de diseminación recoge información de tres tipos: topología de la red, estado de los enlaces y estado de los sistemas finales. Se incluye información de los sistemas finales incluyendo la lista de servicios que ofrece (p.e. servidor DHCP). Desde el punto de vista de conmutación, se obtienen las ventajas del encaminamiento: eficiente utilización de infraestructura, caminos múltiples, ingeniería de tráfico, etc. Los principales inconvenientes de esta propuesta son la incompatibilidad con las redes Ethernet existentes el impacto en los sistemas finales para ser compatibles y que la escalabilidad para redes muy grandes es discutible.

En cuanto a la segunda alternativa, de Plano de control distribuido (DCP) funciona mediante primitivas emitidas por los sistemas finales denominadas: Registrar, Estado e Interrogación. Los sistemas finales se registran en su Bridge local y todos los bridges entre sí intercambian su información de sistemas finales y conectividad mediante los mensajes de Estado. Los mensajes de Interrogación se utilizan para el descubrimiento de servicios por parte de los sistemas finales. Para poder escalar hasta un millón de nodos (miles de bridges) se propone el uso de algoritmos de camino más corto incrementales [AJY00] en vez del algoritmo de Djikstra. Esta alternativa tiene seguramente su punto débil en el gran intercambio de información de sistemas finales entre los bridges. La sobrecarga de ancho de banda por intercambio de datos de sistemas finales es algo menor de 2 Mbps por enlace en el caso estudiado, reduciéndose bajo supuestos de red estable.

2.5.4 Encaminamiento distribuido: Smartbridge

Smartbridge [RTA00], al igual que Autonet, tiene como objetivo lograr bridges capaces de encaminar por caminos mínimos. El encaminamiento tiene dos partes: encaminar entre LANs y llegar a los sistemas finales. Encaminar entre LANs es relativamente fácil, pero es también necesario conocer a qué LAN está cada sistema final conectado, dado que los sistemas finales se pueden mover. Smartbridge utiliza el concepto de *source tree* (unión de las rutas que parten de una determinada LAN) para detectar los sistemas finales que se han movido y el de *destination tree* (unión de las rutas que llegan a una LAN) para implementar el reenvío *(forwarding)*.

Entre los requisitos de diseño de Smartbridge, es de notar que se especifica que la trama no debe ser alterada por el bridge. Por este motivo se justifica la no utilización de encaminamiento de estado de enlaces al no poder disponer de los mecanismos antibucle (como el campo TTL de los paquetes IP) necesarios mientras el protocolo converge.

Smartbridge, a diferencia de Autonet, admite conexiones con segmento compartido e incorpora, como Autonet, el concepto de *diffusing computation* [DS80]. Este mecanismo permite separar las partes de la red que están en reconfiguración del resto. En este mecanismo, el iniciador hace peticiones a sus vecinos, los cuales las realizan a los suyos y así sucesivamente. La respuesta se envía al peticionario cuando el requerido ha contestado, por eso cuando el iniciador recibe la respuesta de todos sus vecinos conoce que se ha completado el proceso. Para asegurar la consistencia, los cálculos deben separarse, de forma que no se mezclen procesos con información antigua y nueva. Esta separación se denomina *wavefront.*

Los procesos básicos de *Smartbridge* son los de *elaboración de inventario* y *aprendizaje de la topología*. El primero consiste en la elección de puerto designado del segmento LAN encargado de enviar todas las tramas. El segundo *(topology acquisition)* es una *diffusing computation* que se propaga por todos los bridges recopilando un lista de todas las conexiones de bridge a segmento y distribuye esta lista (que es una descripción de la topología) a todos los bridges. Cada instancia de obtención de topología se distingue por la UID del bridge iniciador y una marca de tiempo. Existen varias instancias simultáneas de adquisición de topología que se propagan por los bridges compitiendo. La última que finaliza contiene la topología actual.

Smartbridge no es compatible con los bridges 802.1D, aunque puede utilizar hardware estándar de PCs y tarjetas de red implementándolo en software. Encamina por caminos mínimos y requiere mínima configuración. Entre los inconvenientes destacan el no incluir la velocidad del enlace como criterio de elección de caminos y el utilizar reinicialización global ante cambios de topología (*global reboot*). Parece deseable un mecanismo más local y de menor impacto para reconfigurar en caso de cambios de topología.

2.5.5 Encaminamiento distribuido: RBridges

Rbridges es el término adoptado para el nombre de Routing Bridges Transparentes. Según la propuesta de R. Perlman, por Routing Bridge Transparente [PW03] se entiende un dispositivo híbrido que combina las ventajas de los bridges y de los routers. Al igual que en otras propuestas como STAR, el bridge encapsula y reenvía la trama recibida del sistema final unido al segmento LAN del que el bridge designado es único agente.

La discusión sobre RBridges se ha realizado en el IETF en [Rbridge], habiéndose aprobado en Junio de 2005 la formación de un grupo de trabajo (WG) denominado TRILL (Transparent Interconnection of Lots of Links) [PTY04]. Las redes objetivo son redes campus de dimensión no especificada (similar a las actuales, no se contemplan redes grandes como objetivo), con enlaces que comparten un único prefijo de segmento (o subred) IP, utilizan encaminamiento mediante direcciones MAC con protocolo de Estado de Enlaces, encapsulado adicional en capa 2 y un campo adicional de tipo TTL para prevención de bucles. En los Rbridges se intenta combinar las ventajas de los switches (autoconfiguración, transparencia) con las de los routers (prevención de tormentas de tramas, utilización eficaz de la infraestructura,..) evitando los inconvenientes de los bridges (topología activa limitada al árbol de expansión, ausencia de campo TTL, convergencia lenta) y de los routers (las direcciones IP de los sistemas finales están asociadas a cada enlace y cambian al desplazarse el sistema final de sitio en la red, los routers precisan configurar con un prefijo cada enlace, el rango de direcciones IP no es utilizado completamente debido a la fragmentación en subredes).

El objetivo del grupo de trabajo es el diseño de un dispositivo híbrido de router y bridge que combine la sencillez de configuración y el aprovechamiento de la infraestructura instalada. Los requisitos que han establecido son: autoconfiguración de los RBridges, independencia de la dirección IP de los sistemas finales de su lugar de conexión en la red, encaminamiento por caminos mínimos, utilización de los enlaces redundantes, evitar o reducir la difusión en toda la red de los paquetes ARP y de descubrimiento de vecinos (*Neighbor Discovery (ND))*, soportar el descubrimiento de seguro de vecinos *(Secure Neighbor Discovery)*, campo TTL en la trama para estabilidad ante bucles temporales, compatibilidad IPv4 e IPv6, compatibilidad con multicast y seguridad de nivel no peor que el de los bridges actuales.

El protocolo de encaminamiento que por el momento se prefiere es el protocolo IS-IS. Será objetivo del WG el especificar el protocolo, extendido (caso de ser IS-IS) para incluir la alcanzabilidad de los sistemas finales con direcciones de capa dos, soportar difusión y admitir configuración nula. Otros objetivos son la interconexión de distintas tecnologías de capa dos y el soportar protocolos no IP, aunque seguramente no ambos simultáneamente. En la propuesta original se contempla la figura de *RBridge designado (DR)*, el bridge responsable de enviar todas las tramas de un segmento LAN, de forma similar a la empleada en IS-IS. A fin de evitar la interferencia entre capas de protocolo, los aspectos de capa tres a abordar se limitan a los mencionados de limitación del tráfico broadcast y multicast y las funciones de capa tres necesarias para interconexión de tecnologías distintas de capa dos.

Los resultados esperados del WG son, además de la especificación de requisitos del protocolo de encaminamiento y la selección de un protocolo recomendado, los siguientes: la definición de la cabecera de encapsulado de los paquetes, la definición de cómo

realizar ARP y ND así como la definición de la interconexión de tecnologías cuando las direcciones de capa dos tienen distintas longitudes. Y finalmente un análisis de vulnerabilidad y posibles ataques a los dispositivos. Se contempla la coordinación del grupo de trabajo TRILL con el de L2VPN para las soluciones a los temas comunes (como ND y ARP). Con el IEEE solamente se contempla un enlace, sin coordinación de grupos.

Los objetivos de TRILL coinciden en buena parte con los establecidos en nuestra definición del problema, siendo las principales diferencias nuestro objetivo de redes campus grandes (objetivo excluido de TRILL) y el circunscribirnos a redes Ethernet. El enfoque inicial de las soluciones es divergente, en parte porque en el escenario de redes campus muy grandes el encaminamiento basado en direcciones MAC produce tablas bastante grandes porque las direcciones MAC por ser planas no permiten agregar rutas en el encaminamiento, con lo que la escalabilidad puede quedar limitada. Actualmente la propuesta contempla como posibilidad opciones de inspección del paquete y de conmutación basada en las direcciones IP aprendidas dentro de la trama. Dadas las dimensiones de las redes campus objeto de nuestro problema, las propuestas actuales de este grupo no parecen escalar económicamente a una red de 100.000 terminales, requiriendo excesiva memoria en los bridges. Otra razón de la divergencia es que, mientras que en algoritmos de encaminamiento en redes fijas parece haber pocos aspectos susceptibles de mejora significativa, no ocurre así en cuanto a los árboles múltiples de expansión, terreno escasamente explorado en la práctica fuera del estándar MSTP y susceptible de optimización.

2.5.6 Encaminamiento distribuido: Link State Over MAC (LSOM)

LSOM [GDS03] utiliza el mismo encaminamiento basado en direcciones MAC que RBridges, pero en troncales ópticos Ethernet de área metropolitana terminados en routers en vez de en redes campus. Es un protocolo de encaminamiento de estado de enlace sobre direcciones MAC. En la Figura 8 se muestra un ejemplo de red objetivo de LSOM.

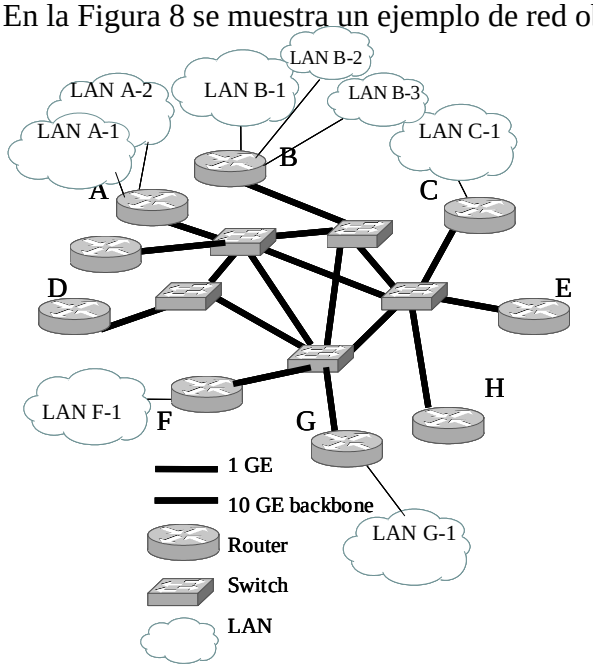


Figura 8. Red Metropolitana delimitada con routers.

Esta red consta de una serie de conmutadores 10 GE Ethernet interconectados entre sí formando una red troncal y a las respectivas LAN mediante los correspondientes encaminadores. Estos routers dividen la red en subredes IP. La solución se ajusta a este escenario debido ael reducido número de direcciones MAC que el protocolo debe manejar (solamente los interfaces de troncal de los routers) lo hace indicado. Dentro de LSOM se han estudiado y comparado las capacidades de troncales basados en árbol de expansión, en Resilient Packet Ring (RPR) y LSOM. LSOM utiliza caminos óptimos, distribuye el tráfico en la red y utiliza bien la infraestructura, pero no resuelve el problema de la proliferación de direcciones MAC.

2.6 Combinaciones de bridge y router.

Se analizan en esta sección los dispositivos que incorporan funciones de bridge y router que no se han descrito en secciones anteriores.

2.6.1 Brouter.

Lin y Gerla utilizaron el término *Brouter* para el dispositivo híbrido de bridge y router que propusieron en [TG91]. Los objetivos eran combinar la transparencia para los sistemas finales, caminos óptimos y la reducción de los tamaños de tabla de encamiento en los bridges. El protocolo Brouter se compone de cuatro mecanismos principales: determinación de identidades de las LAN, mecanismo de autoaprendizaje y encaminamiento, intercambio de vectores de mapeado e intercambio de tablas de retardos en el algoritmo distribuido de encaminamiento. Se utiliza una tabla de correspondencia entre ID de LAN y sistemas finales conectados a cada LAN. La Tabla de encaminamiento contiene: (LAN ID destino, retardo a la LAN, puerto de salida, siguiente Brouter) y Tabla de Vecinos (Puerto, ID LAN adyacente, Brouters adyacentes) . La elección del Brouter designado de cada LAN se realiza por menor Brouter ID y la identidad de la LAN se forma concatenando la identidad del brouter designado y la identidad de puerto conectado a la LAN. La propuesta no es compatible con los bridges 802.1D y tiene una sobrecarga importante en el proceso para obtener las tablas de retardos por puerto.

2.6.2 Conmutadores de capa tres.

A finales de los años noventa, con el desarrollo de los conmutadores Ethernet, las redes campus aumentaron su capacidad, creando un cuello de botella en las rede troncales de las redes empresariales y con ello la necesidad de encaminadores de prestaciones similares a la de los conmutadores en el nivel de Distribución para separar y encaminar entre los segmentos IP de la red campus. A ello contribuyó el aumento del tráfico en Internet con el éxito de HTTP. Dado que los conmutadores se implementaban en hardware y los encaminadores realizaban el encaminamiento por software en la CPU, las diferencias de costes y prestaciones eran muy importantes. Esta búsqueda de encaminadores que tuvieran las prestaciones de los conmutadores y precios menores que

los encaminadores produjo la aparición desde 1997[1] [Anr98] de dispositivos de este tipo denominados de diversas formas: *routing switch, wirespeed router, layer 3 switch, IP switch, routing switch, switching router* en implementaciones propietarias, pero algunos compatibles con los protocolos estándar.

El término que tiende a predominar es el de conmutador multicapa (*multilayer switch),* el cual incluye los conmutadores que tienen en cuenta capas superiores a la de red, como la de transporte, para realizar la conmutación. Funcionalmente estos dispositivos engloban un encaminador y un conmutador conectados en paralelo como se muestra en la Figura 9. Los tipos de conmutadores multicapa más destacables son los del tipo *Route Once-Switch Many* en los que se calcula la ruta en el primer paquete y los paquetes sucesivos al mismo destino son conmutados en vez de encaminados utilizando diversas tecnologías propietarias. Los *wirespeed routers* y los *switching routers* son auténticos routers no propietarios y basados en ASICs que realizan el encaminamiento de capa 3 en hardware e implementan OSPF o RIP para comunicarse con el encaminador WAN, dado que carecen de capacidad de encaminamiento WAN. Estos dispositivos determinan la ruta basados en la información de capa 3, verifican la integridad de la cabecera de capa 3, la caducidad del paquete, procesan las opciones del paquete IP y actualizan las estadísticas de la MIB. Al estar implementados en hardware, permiten también alto rendimiento en la aplicación de políticas. Finalmente los denominados *Layer 3 Learning Bridging ó routing switches* [CH01], que no realizan encaminamiento sino que inspeccionan los paquetes IP y realizan un aprendizaje de las relaciones de las direcciones IP/MAC de los sistemas finales que les permite conmutar las tramas en vez de encaminarlas. Los de este tipo no soportan protocolos dinámicos como RIP u OSPF, son propietarios, y no modifican el paquete IP (decremento de TTL y cambio de direcciones MAC origen y destino).

Aunque los conmutadores multicapa son más simples de configurar que los encaminadores, requieren la correspondiente configuración, consistiendo principalmente en la definición de las subredes IP. Debe mantenerse la integridad de las subredes. Los conmutadores de capa tres tienen dominios implícitos de conmutación en cada interfaz, por lo que pueden agruparse puertos con el mismo dominio de conmutación, lo que permite agregar capacidad de forma sencilla, mientras que en un encaminador tradicional la subred se define a nivel de puerto individual.

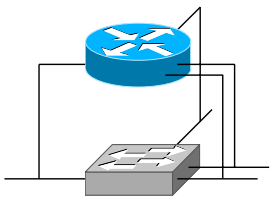


Figura 9. Esquema conceptual de conmutador de capa tres.

[1] La integración de funciones de switch y router en 3com comienza en 1992, pero con implementación software.

2.7 Registro e Inspección.

El mecanismo básico en los bridges transparentes para conocimiento de los sistemas finales es el de aprendizaje en cada puerto de los sistemas finales conectados a través de la LAN correspondiente. Pero entre los protocolos estándar de capa dos existen algunos que pueden ser utilizados para registrar los sistemas finales en los bridges. Ejemplos de estos protocolos son GMRP (Generic Multicast Registration Protocol) y GVRP (Generic VLAN Registration Protocol) , respectivamente para registro multicast y para registro en VLAN por parte de los sistemas finales y difusión de dicho registro entre los bridges de la red. Ambos protocolos utilizan como soporte el protocolo GARP (Generic Attribute Registration Protocol) [802.1D04]. La propuesta descrita en [HT03] está orientada a soportar la movilidad de los usuarios de Internet en los switches y consiste en combinar el protocolo estándar GARP con el mecanismo de aprendizaje de los Bridges transparentes. Se definen como mecanismos alternativos en los Bridges los de Registro (REG) del host e Inspección (SNOOP) de trama. El primero usaría GARP y el segundo es el aprendizaje estándar de direcciones MAC de los Bridges. Se trata de mejorar el soporte de la movilidad de sistemas finales en grandes LAN.

El mecanismo habitual de aprendizaje (denominado aquí SNOOP) de los Bridges pierde su efectividad y se sobrecarga cuando los sistemas finales tienen gran movilidad consumiendo mucho ancho de banda por inundación al no actualizarse las tablas adecuadamente, pudiendo haber pérdida de paquetes. Puede utilizarse alternativamente el protocolo GARP para registrar los sistemas finales en lugar de aprenderlos, pero esto generaría registros de gran tamaño en los Bridges. Se propone entonces una solución combinada (REG-SNOOP). Para evitar el registro del sistema final en todos los Bridges, el sistema final se registra solamente en un subconjunto de los Bridges, los más cercanos. En el resto de los casos los Bridges hacen difusión hacia el sistema final destino y utilizan aprendizaje. Se han realizado simulaciones que visualizan con claridad los compromisos de diseño entre el radio de la zona de Bridges para registro, la vida media de los registros y el consumo de ancho de banda. Igualmente se han realizado simulaciones para estimar el tamaño medio de la base de datos.

Los inconvenientes apreciados son, por un lado la necesidad de optimizar el sistema adaptando los parámetros mencionados al tráfico y movilidad esperados según las aplicaciones, impredecibles a priori en el caso de redes campus y por otro la sobrecarga en los Bridges debida a los registros GARP. El enfoque de análisis utilizado parece válido para valorar otras posibles soluciones mixtas parecidas. Actualmente se prepara un nuevo estándar IEEE 802.1ak [8021ak] para un Protocolo de Registro Múltiple (Multiple Registration Protocol (MRP)) que posibilite un registro de atributos más eficiente que GARP. Esto podría suponer una mejora significativa de las prestaciones de la propuesta analizada si se sustituyera el protocolo de registro por MRP.

2.8 Bridging Jerárquico.

En esta sección se ilustra el problema de escalabilidad de las redes conmutadas y se agrupan las tecnologías dentro de Ethernet que utilizan jerarquía para abordar el problema del tamaño máximo de dominio conmutado o de comunicación entre dominios.

2.8.1 Limitaciones escalabilidad VLANs.

La etiqueta para Redes Locales Virtuales (VLAN), especificada en el estándar 802.1Q, fue la primera extensión de la trama Ethernet aprobada por el IEEE. El empleo de VLANs permite la coexistencia simultánea e independiente en una infraestructura común de dominios de difusión separados, los cuales requieren de un router para interconectarse. Pero esta extensión de la trama presenta limitaciones importantes para obtener escalabilidad. La principal limitación de 802.1Q cara a la escalabilidad de las redes Ethernet se encuentra en que el número de 4094 VLANs puede ser pronto insuficiente para un proveedor que ofrezca este tipo de servicios. [HAD01] es una propuesta para Ethernet en redes metropolitanas que utiliza doble encapsulado Ethernet para posibilitar la escalabilidad. El encapsulado adicional de capa dos en Ethernet realizado mediante etiquetas adicionales previene el problema de explosión de direcciones MAC en la red metropolitana y separa los dominios de difusión. Dado el interés existente por extender Ethernet a ámbitos de red cada vez mayores, el *etiquetado* Ethernet, (incluyendo en el término etiquetado a todo tipo de etiquetas (*tag)* adicionales a la trama básica Ethernet), es tema de gran interés dentro del IEEE dada su relación con la extensión de diversas funcionalidades esenciales en las redes MAN y VPNs de capa 2 (L2VPN). Los esquemas de encapsulado Ethernet principales son actualmente los siguientes [CHI+02]: Apilado de VLANs (tambien llamado *Q-in-Q*), el denominado VMAN y finalmente Ethernet sobre MPLS. En los siguientes epígrafes se describe un precedente de encapsulado y la situación presente en lo relativo al encapsulado Ethernet más relacionado con el problema.

2.8.2 Virtual Bridged LAN

La primera propuesta de bridges virtuales o de jerarquización de bridges de la que tenemos noticias es la de Casale, Catania y otros [CPV93] [CAS+94]. Para unir LANs de forma escalable y fiable se propone consideran una serie de islas de redes LAN que se interconectan unas con otras mediante bridges remotos. En la Figura 10 se muestra la disposición de las islas de LANs. En cada isla se ejecuta un algoritmo de árbol (único) de expansión independiente de los demás. El bridge virtual (*virtual root*) está formado por el conjunto los bridges remotos interconectados. Estos bridges remotos ejecutan un protocolo de árbol expansión independiente que opera en un nivel jerárquico superior. Esta separación entre los árboles de expansión central y de isla se obtiene utilizando BPDUs de dos tipos diferentes, una para cada nivel jerárquico, y mediante encapsulado de las tramas cuando se envían desde una isla hacia el bridge virtual.

Este diseño de red permite obtener mucha mayor fiabilidad y escalabilidad que la que se obtendría si se interconectaran mediante un único Árbol de Expansión. Recordemos la

limitación establecida por las recomendaciones (no imperativas) del estándar 802.1D para el diámetro máximo de la red (7 saltos). La literatura incluye el modelo para el cálculo de fiabilidad resultante.

Dentro de la perspectiva de la Tesis, se consideran aplicables estos conceptos a las redes campus de nuestro estudio por el desacoplo que efectúan entre las redes constituyentes mejorando la escalabilidad del árbol de expansión, pero se consideran necesarias alternativas al árbol de expansión que realicen una mejor aprovechamiento de la red y utilicen caminos más cortos.

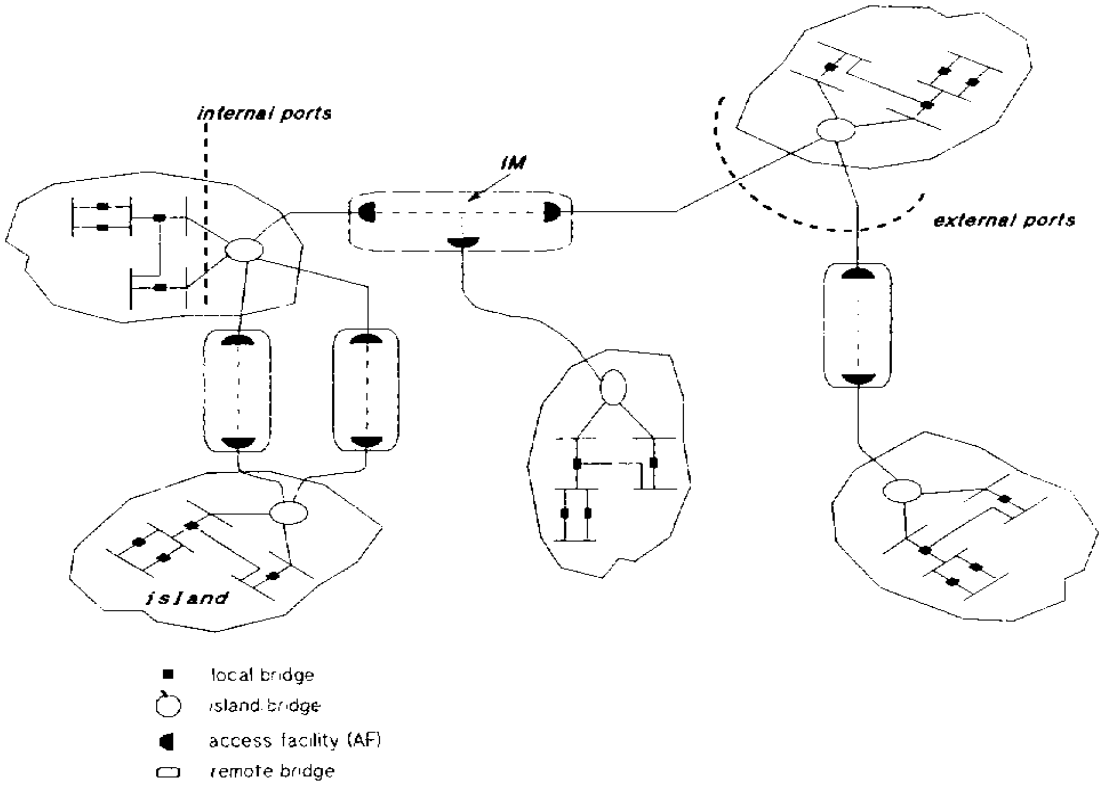


Figura 10. Bridge Virtual para interconexión de islas de LANs (Casale et al)

2.8.3 Redes y servicios Metropolitanos Ethernet.

En esta sección solamente mencionamos las propuestas relacionadas con Ethernet en entornos metropolitanos que consideramos de interés para el problema descrito. En las secciones anteriores se han descrito algunas propuestas, provenientes del área de las redes metropolitanas Ethernet como LSOM y del área de Redes Privadas Virtuales de Capa 2 (L2VPN) como Global Open Ethernet. Se han mencionado en la medida en que algunos de los problemas son comunes a las redes campus de gran tamaño objeto de esta Tesis. Bajo la hipótesis de trabajo de que el troncal de las redes campus, al aumentar éstas de tamaño, se asemeja a un troncal de red metropolitana, pueden ser aplicables soluciones propuestas para redes troncales MAN en la medida en que lo sean al troncal de redes campus de gran tamaño. También son de interés a efectos de interoperabilidad en capa dos entre redes campus y metropolitanas. Los puntos en común con el problema planteado en esta Tesis son los problemas que resultan de ser redes Ethernet de gran tamaño, es decir:

-Insuficiencias del encaminamiento en capa dos. Poca adecuación, convergencia lenta y bajo rendimiento del Árbol de Expansión.

-Inundación de difusiones (*broadcasts)* y necesidad de compartimentación de la red.

-Proliferación de direcciones MAC (*MAC explosion problem)* y falta de agregabilidad de las direcciones MAC por ser planas, no jerárquicas.

Respecto a las diferencias con las redes campus, la principal radica en las autoridades de los dominios. Las redes campus se suponen bajo una autoridad de gestión única y sin intervención de proveedores de redes como es el caso de las redes Ethernet metropolitanas. Las redes Ethernet metropolitanas se orientan a proporcionar, mediante una red propiedad de un proveedor y administrada por dicho proveedor, la interconexión lo más transparente posible de diversas redes campus propiedad de una mismo cliente. Esto otorga un papel central a la gestión, control y administración de dichas redes, en dirección divergente respecto a la autoconfiguración buscada en las redes campus. Estas diferencias reducen la aplicabilidad de las soluciones a redes campus grandes, si bien la tipología de las soluciones puede ser cercana en algunos aspectos como lo es el uso de encapsulado adicional para evitar la proliferación de direcciones MAC.

La economía de Ethernet hace que tienda a extenderse a todos los ámbitos hasta ser una tecnología extremo a extremo. A medida que Ethernet se va convirtiendo en la interfaz de usuario-red (User-Network Interface (UNI)) predominante para acceder a las MAN y WAN, las organizaciones de estandarización como el IETF, UIT, IEEE y el Metro Ethernet Forum están definiendo servicios Ethernet sobre redes WAN y MAN [ZFB04] [Chi+02][Chi+04]. Solamente son cercanas al problema abordado en esta Tesis las propuestas basadas en extensión de Ethernet, no las basadas en MPLS o túneles de otro tipo. El grupo IEEE 802.1 es el responsable de elaborar los estándares para redes MAN. La trama Ethernet está siendo extendida con etiquetas adicionales diversas según la funcionalidad introducida. La extensión de Ethernet para los proveedores de servicios se realiza dentro del grupo de trabajo Provider Bridges IEEE 802.1ad [802.1ad], que se elabora como enmienda a IEEE 802.1Q con encapsulado Q-in-Q. Se distingue entre VLANs de cliente (C-VLAN) y VLANs de servicio (S-VLAN) y las correspondientes funcionalidades en los bridges de la red del proveedor que las soporten. Numerosas implementaciones de fabricantes han hecho posible el despliegue de equipos. El grupo de Provider Backbone Bridges 802.1ah trabaja a un nivel superior (troncal) para complementar a 802.1ad, posibilitar la interoperabilidad entre redes de distintos proveedores y facilitar la escalabilidad e interconexión de S-VLANs.

2.8.3.1 Resilient Packet Ring (RPR).

El estándar Resilient Packet Ring, promovido en 2001 y aprobado como 802.17 [RPR] en 2004 trata de responder con tecnología Ethernet a las carencias de SONET/SDH y a su obsolescencia en las redes en anillos de fibra óptica. Las principales carencias de SONET/SDH son su elevado coste, la ineficiencia para el tráfico de paquetes de la asignación en modo circuito de ancho de banda estático. Igualmente ineficiente es el ancho de banda reservado para protección de los circuitos establecidos y el provisionamiento del servicio es lento. Estas carencias empujan hacia la sustitución de los anillos SDH. Pero el estándar Ethernet no satisface los requisitos de SDH porque el árbol de expansión no soporta las topologías de anillos múltiples requeridas y no tiene la rapidez requerida por el estándar de protección de SDH (reconfiguración en 50 mseg.). Asimismo es necesaria una separación de cliente/servicio en capa dos que sea escalable.

RPR tiene como objetivo ser alternativa a SDH mediante tecnología Ethernet resolviendo dichas limitaciones.

RPR utiliza paquetes de longitud variable que utilizan direcciones MAC 802 con direcciones origen y destino como Ethernet. Funcionalidades específicas de RPR que lo diferencian de Ethernet, además de la reconfiguración, son la de *drop and continue*, los paquetes broadcast y multicast se extraen del anillo y a la vez se reenvían por el anillo. Soporta Clase de Servicio en la cabecera para manejar prioridades de forma compatible con Ethernet y utiliza campo TTL para evitar bucles en el anillo. RPR utiliza anillos que cursan tráfico en sentidos opuestos a fin de que los nodos tengan rutas alternativas de comunicación entre ellos en caso de fallo, no utiliza reserva de ancho de banda en modo circuito, no utiliza nodo maestro de anillo y es autoconfigurable. El estándar incluye enmiendas a 802.1D para la compatibilidad con Ethernet.

RPR está orientado a topologías en anillo, lo que limita mucho su aplicabilidad, haciendo inadecuada su aplicación en redes campus (malladas) en el problema que nos ocupa, dado el bajo rendimiento (longitud de camino y rápida saturación de la red) que presentan las topologías en anillo [GDS03].

2.9 Arquitectura y evolución de las redes campus

Tras revisar el estado de los protocolos, y para tener cierta perspectiva histórica, se resumen brevemente las tipologías de redes campus y su evolución en el tiempo:

Las primeras redes Ethernet tuvieron una arquitectura llamada de "*Router y Concentradores*". Un router se conectaba a una serie de concentradores que compartían el medio Ethernet para acceder a dicho router. En esta disposición, cada concentrador supone un cuello de botella para el tráfico por compartirse el medio físico entre todos los enlaces conectados al concentrador. El router también constituye otro cuello de botella para el tráfico de entrada y salida.

Este diseño quedó obsoleto tras la aparición y difusión de los conmutadores (*switches*), bridges transparentes multipuerto, que eliminan los dominios de colisión de los concentradores. El éxito de los bridges transparentes multipuerto o conmutadores (*switches*) produjo la idea errónea de que la topología ideal para redes campus consistía en redes planas, sin routers, exclusivamente conmutadas y configurando VLANs de extremo a extremo de la red campus. Este tipo de redes tiene una gran capacidad pero son muy complejas de administrar, carecen de estructura lógica y tienen dominios de árboles de expansión que se solapan. Esta complejidad resultante hace a veces necesario limitar la redundancia para obtener la estabilidad de la red. Además estas redes no escalan y violan la regla de diseño de una VLAN por subred IP.

La experiencia y evolución posteriores han llevado al modelo actual de Redes Multicapa mostrado en la Figura 11 cuyas funciones se esquematizan en la Figura 12.

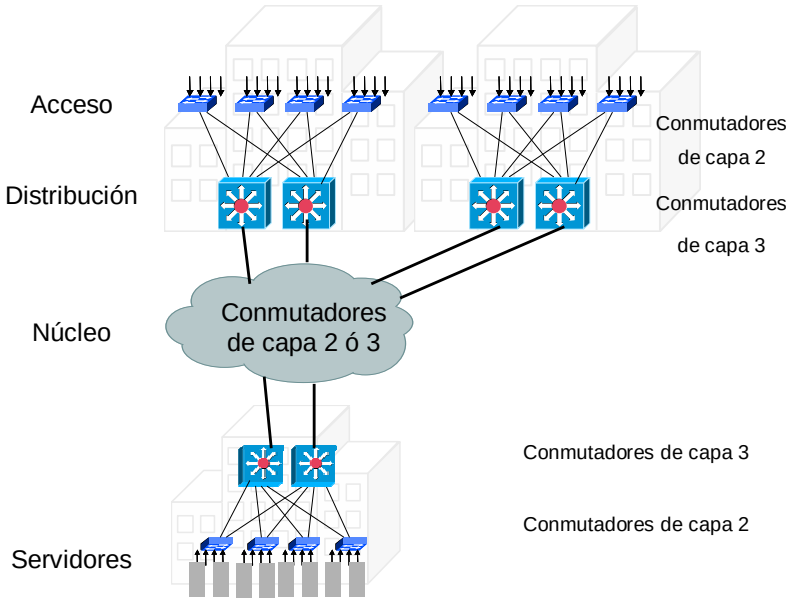


Figura 11. Esquema de red campus multicapa.

En este modelo se distinguen tres capas, cada una con unas funciones principales y unos requisitos que se derivan de dichas funciones. Las capas se denominan como capa de Núcleo o Troncal, capa de Distribución y capa de Acceso. El Núcleo requiere altas prestaciones, por lo que sus funcionalidades deben ser solamente las esenciales. Puede utilizarse conmutación en capa 2 o en capa 3. Tiene como principal y exclusiva funcionalidad la conmutación a alta velocidad, sin requisitos especiales de filtrado o procesado que reduzcan las prestaciones. El núcleo debe ser predecible, para lo cual debe tener un comportamiento determinista y muy fiable.

La capa de Distribución debe compartimentar la red en subredes para hacerla más robusta, por lo que suelen emplearse routers o conmutadores multicapa, que permiten dividirla en segmentos o subredes IP confinando los posibles problemas de una subred conmutada al ámbito de esa subred. La capa de Distribución implementa las políticas de conectividad, qué usuarios se conectan a qué y a través de qué enlaces. En la capa de Distribución se realiza el encaminamiento entre VLANs, la seguridad, la agregación de direcciones, el filtrado y otras muchas funciones.

Finalmente la capa de Acceso debe optimizarse para un acceso económico y controlado a la red, por lo que las funcionalidades y requisitos se orientarán a ello: gran densidad de puertos por cada conmutador, economía y funciones de control de acceso. La capa de Acceso facilita el acceso tanto local como remoto de los usuarios. En las capas de Distribución y Acceso se recomiendan a veces topologías antibucle para obtener convergencia más segura y rápida. Es prioritario optimizar la utilización de la infraestructura más costosa, como lo es el equipo y la fibra óptica de los enlaces del Núcleo y de Distribución (enlaces de 1 Gigabit y 10 Gigabit).

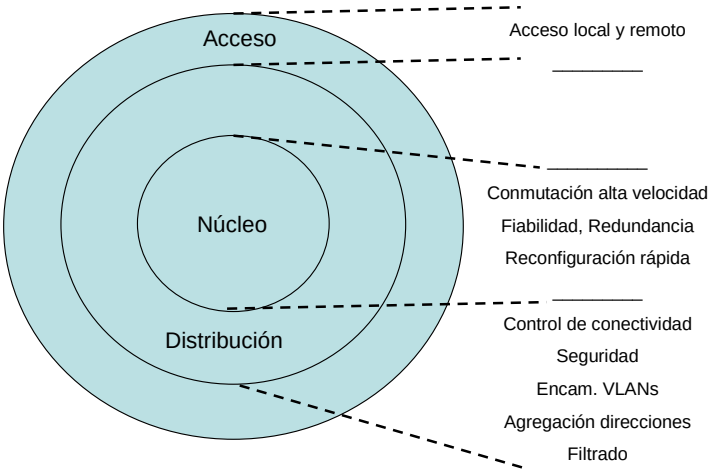


Figura 12. Funciones en las redes campus multicapa.

En las redes campus reales suelen coexistir diversas tecnologías, resultado de la instalación de equipos para modernización y ampliación de la red a lo largo del tiempo. En la actualidad los equipos instalados son principalmente Ethernet y ATM. Los conmutadores ATM, muy difundidos en el pasado por su alta capacidad, han sido generalmente sustituidos por conmutadores Gigabit Ethernet, un ejemplo de ello es la red de la Universidad Carlos III. Ciñéndonos a las tecnologías Ethernet, coexisten actualmente cuatro generaciones de Ethernet en las redes campus: 10 Mbps, 100 Mbps, 1 GE y 10 GE. En cuanto a su ciclo de vida, Fast Ethernet (100 Mbps) se encuentra en fase de madurez con precios banalizados, Gigabit Ethernet acercándose a la madurez y 10 Gigabit Ethernet en fase de crecimiento rápido inicial tras su estandarización en 2002.

Los costes económicos de los conmutadores Ethernet dependen básicamente de su velocidad y prestaciones adicionales. El coste de los enlaces Ethernet depende del medio de transmisión (cobre o fibra óptica). El medio de transmisión se elige según la distancia entre nodos: dadas las limitaciones de los enlaces de cobre, el cobre es excluido a partir de la distancia límite de la tecnología utilizada. Por ello lo habitual es que los troncales no colapsados utilicen fibra óptica. Los enlaces de distribución también. Un enlace adicional de fibra de troncal será mucho más costoso que uno de Acceso tanto por el uso de fibra como por el coste del equipo por puerto con una componente de coste fija y otra variable, función ésta última de la longitud del enlace.

2.9.1 Estado del arte en conmutadores

Los bridges multipuerto, comercialmente denominados conmutadores (*switches)*, han evolucionado de forma paralela a las velocidades de transmisión estandarizadas en Ethernet: 10 Mbps, 100 Mbps, 1 Gbps, 10 Gbps. En sus orígenes eran dispositivos basados en software, pasando posteriormente la funcionalidad al hardware por razones obvias de velocidad y coste. Además del incremento de velocidad, se han sofisticado mucho, dotándose de múltiples funcionalidades como agregación de enlaces, VLANs,

árboles múltiples de expansión, clasificación de paquetes multicapa, listas de admisión (ACL) y otros.

La estructura básica de un switch, en cuanto a su funcionalidad principal de reenvío de tramas entre puertos, cada uno conectado a una LAN (*MAC Relay)*, consiste en unos puertos de entrada con sus colas correspondientes y un conmutador central que interconecta dichos puertos de entrada con los puertos de salida. Esta conmutación puede realizarse de diversas formas : mediante una memoria común donde se escriben las tramas recibidas y de donde serán leídas para transmitirlas al puerto de salida correspondiente, o mediante un bus que interconecta un puerto de entrada con uno o varios de salida en un instante y en otro instante conecta otras parejas de puertos. Así pues, el diseño de los switches tiene dos aspectos principales: la clasificación y filtrado de los paquetes en la entrada y la conmutación de los mismos basada en el resultado de dicha clasificación. El filtrado se realiza según las reglas almacenadas en la base de datos de filtrado en cuanto a admisión de tramas resultando en un encaminamiento diferenciado de las tramas. Se muestra en primer lugar el esquema de funcionamiento de un conmutador, a continuación un ejemplo concreto de conmutador actual y finalmente se pasa revista de forma rápida a los dispositivos de clasificación y procesado de paquetes actuales. En la Figura 13 se muestra el diagrama de bloques genérico de un conmutador.

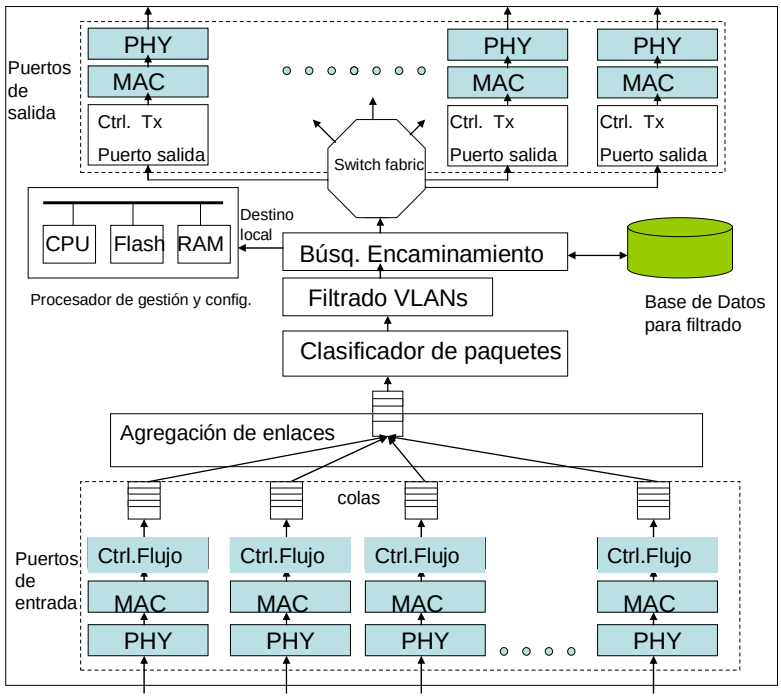


Figura 13. Diagrama de bloques de un conmutador.

El concepto de bridge transparente con autoaprendizaje fue inicialmente propuesto por Kempf [Kem86]. El diseño de los bridges Ethernet ha evolucionado, desde la estructura básica de un dispositivo software que interconectaba dos LAN filtrando las tramas internas de cada LAN, hasta incorporar las funcionalidades añadidas a Ethernet posteriormente, tales como VLANs, agregación de enlaces, control de flujo, etc.

Aunque algunas funciones aparecen conceptualmente como centralizadas, las implementaciones actualmente tienden a distribuirlas por puerto para alcanzar velocidad de conmutación equivalente a la del enlace (*wirespeed)*. Tras las funciones de enlace

físico (PHY) y control de acceso al medio (MAC), se procesan el control de flujo (caso full-duplex) y la agregación de enlaces si está activada. Las tramas que llegan al clasificador son verificadas contra las reglas de admisión del clasificador de paquetes, a continuación se hace el tratamiento por criterios de pertenencia de VLAN y finalmente el encaminamiento basado en la base de datos de filtrado. Las tramas con destino local (tales como las BPDUs de protocolo STP, las de GARP que se terminen o las de gestión del switch), se encaminan al procesador de gestión y configuración.

La Figura 14 muestra un esquema del proceso de filtrado de paquetes y del mantenimiento de la base de datos de filtrado con las altas y bajas en la misma [Sei00]. La salida del clasificador de paquetes opera sobre el buscador (*lookup)*, obteniéndose la dirección de puerto de salida si el resultado es reenvío, el descarte de la trama si no procede su reenvío o su procesamiento local en el procesador del switch.

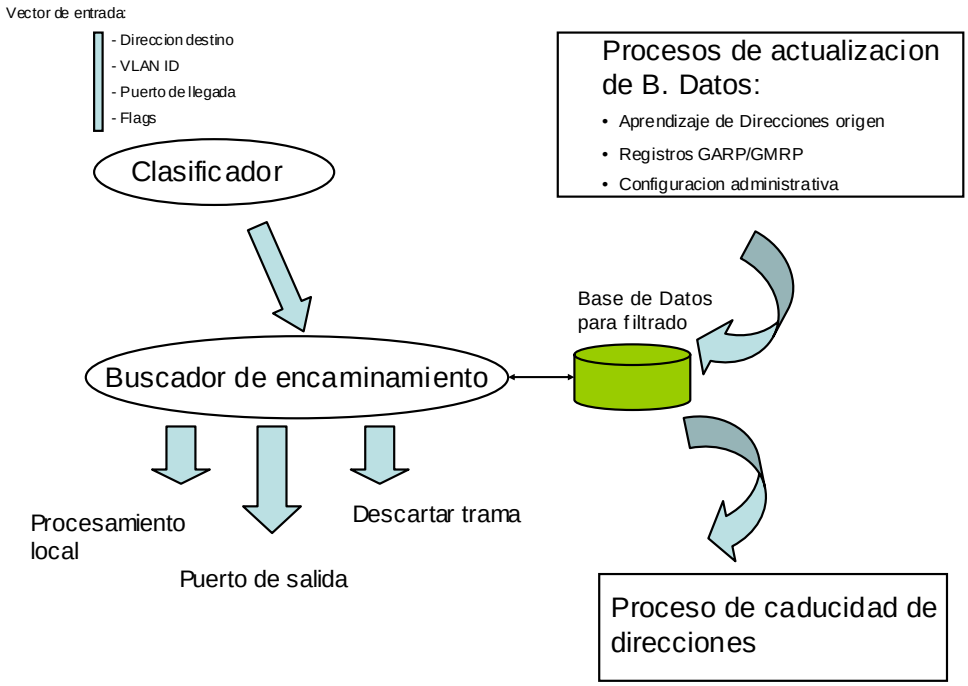


Figura 14. Procesos de filtrado de paquetes en un conmutador.

Como ejemplo del estado del arte de la tecnología en conmutadores, en la Figura 15 se muestra el diagrama de bloques de un conmutador integrado Ethernet de 12 puertos de 10 Gbps que opera con escrituras y lecturas simultáneas y en la Figura 16 su principio operativo de conmutación con escrituras y lecturas simultáneas controladas por un planificador.

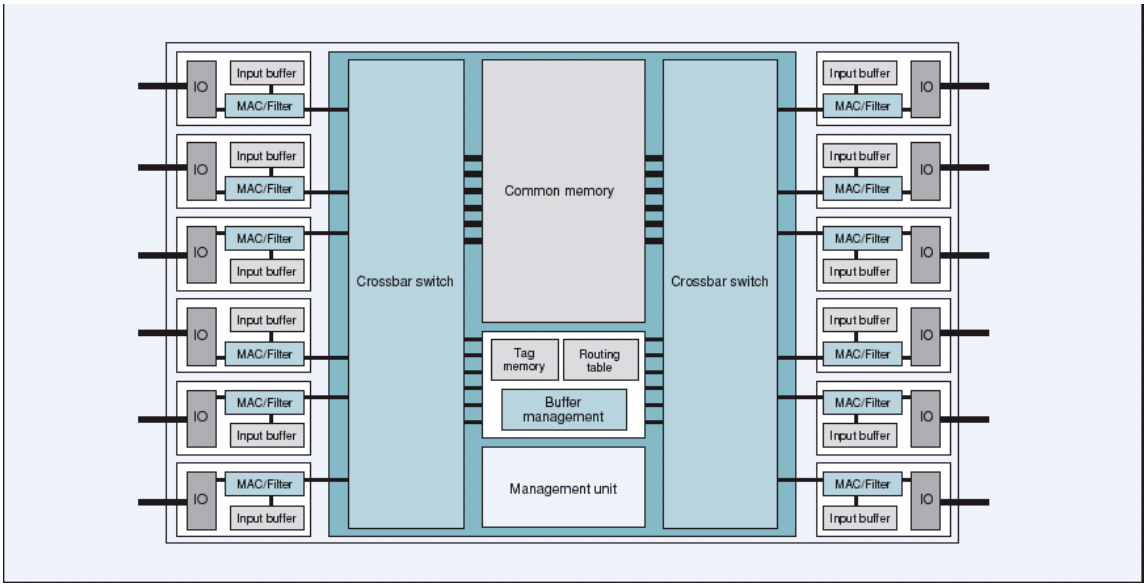


Figura 15. Diagrama de bloques conmutador 12 * 10 Gbps (Fujitsu MB87Q3050)

En cuanto a la clasificación de los paquetes, la disponibilidad de dispositivos de clasificación de paquetes de alta capacidad y rendimiento es cada vez mayor debido a las necesidades de clasificación de paquetes que plantea el uso de Classless Interdomain Routing (CIDR), incrementadas por la necesidad de clasificación multicapa de tráficos, es decir utilizando la información de las capas de aplicación y de transporte del paquete para su clasificación. Estos requisitos pueden suponer en IPv4 un máximo de hasta 120 bits a comparar para clasificar un tráfico y hasta 304 bit en IPv6, a velocidades de enlace de 10 Gb, lo que da idea de la exigencia de dichos requisitos.

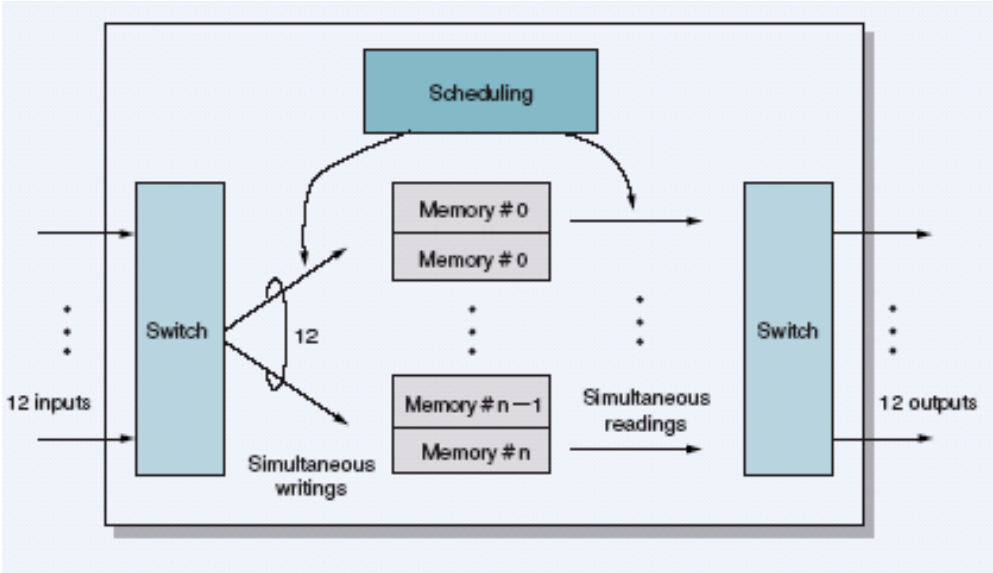


Figura 16. Principio operativo del conmutador MB87Q3050

Los dispositivos de clasificación de paquetes más destacables actualmente son los siguientes:

- Memorias direccionables por contenido (*Content Addressable Memory (CAM)).* Las CAM hacen posible minimizar el tiempo necesario para una búsqueda en tabla *(lookup)* a un solo acceso, dado que se inspecciona toda la memoria en paralelo para comprobar si un determinado contenido existe en la

memoria. La comparación se realiza completa, verificando la coincidencia exacta de "0"s y "1"

- Memorias CAM Ternarias (*Ternary CAM (TCAM)*). Las memorias CAM ternarias presentan una gran ventaja frente a las CAM porque realizan la comparación de contenido de la memoria admitiendo bits, además de en estado 0 y 1, en estado indiferente ("X, *don't care"),* con lo que la obtención del prefijo coincidente más largo (*longest prefix match (LPM))* se realiza en un solo paso con un circuito de arbitrio de prioridades, mientras que con las CAM requiere pasos posteriores de comparación de prefijos. Las TCAM están mas limitadas de tamaño que las CAM, por eso se emplean a veces combinaciones de CAM binarias y ternarias.
- Procesadores de red. (*Network Processors, Packet Processors (NP)).* Los procesadores de red son una de las opciones más utilizadas actualmente. Son procesadores especializados en clasificación de paquetes que pueden utilizar CAMs, TCAMs u otros dispositivos de forma auxiliar conectados externamente o incluidos en el propio chip. Los NP son una alternativa a los ASICs, diseñados específicamente para cada equipo utilizados por los fabricantes para obtener el máximo rendimiento. El diseño de los NP varía con los fabricantes y con la aplicación preferente del NP, si bien se tiende a incorporar el mayor número de funciones posibles excepto las memorias, los interfaces físicos (Ethernet PHY) y el conmutador hardware propiamente dicho *(switch fabric).* En la Figura 17 se muestran el procesador de red X11 y sus interfaces. Este procesador está orientado a la clasificación de paquetes de alta capacidad (compatible IPv6), para lo cual emplea conexiones a TCAM y otras memorias externas. En la Figura 18 se muestra un ejemplo de aplicación en el que puede apreciarse el alto grado de integración de funciones.

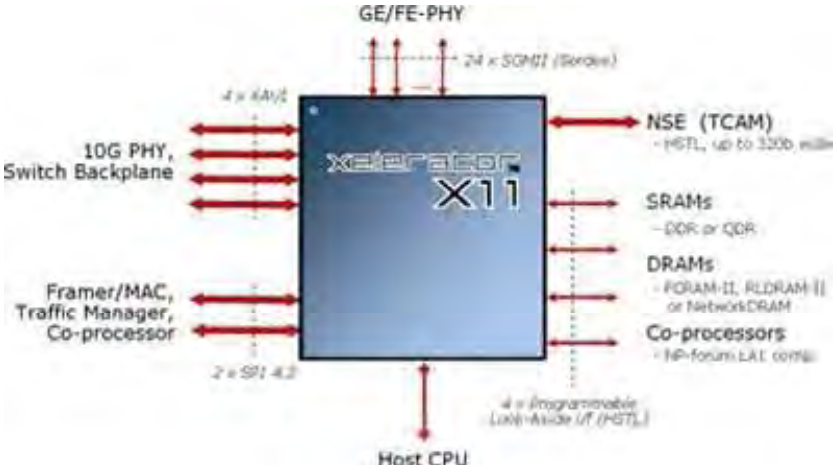


Figura 17. Interfaces de Procesador de Red (Nextel X11).

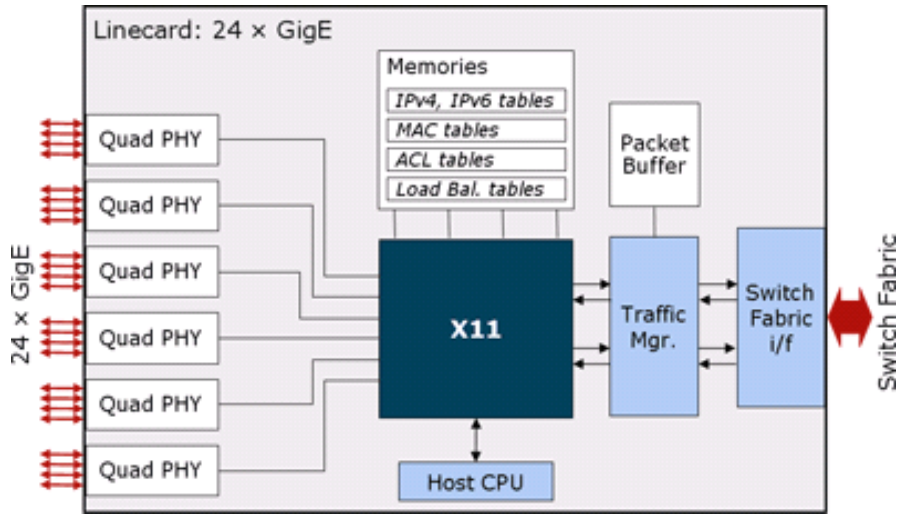


Figura 18. Ejemplo de aplicación de procesador de red en conmutador (Nextel X11)

Dada la situación de la tecnología en cuanto a implementación predominantemente basada en procesadores de red y ASICs, no se ven en principio factores tecnológicamente decisivos que determinen la búsqueda de soluciones en un sentido determinado que facilite su implementación hardware. Un factor tecnológico de importancia como es el tamaño pequeño de las memorias ternarias TCAM no tiene en el caso que nos ocupa importancia dado que la comparación de longitud de longitud máxima de prefijo, tan importante en encaminamiento IP, no es necesaria en encaminamiento por direcciones de capa 2 al no tratarse de direcciones jerárquicas.

2.9.2 Estado actual de conmutadores comerciales

Con el fin de tener una visión del estado de la tecnología en redes campus se han escogido varias fuentes de organizaciones ajenas a suministradores, aunque elaboradas con información procedente de los mismos. En [Wit05] puede verse una muestra actualizada de las capacidades actuales de conmutadores y del estado del arte del diseño de redes campus con conmutadores de 1 Gigabit y opcionalmente 10 Gigabit Ethernet aplicada a una red campus de tamaño medio. Muestra las dos redes alternativas propuestas (una de arquitectura predeterminada, otra elaborada libremente por el suministrador) por diversos fabricantes a unos requisitos determinados de una red campus hipotética. Esta referencia nos da una pauta del estado actual de diseño de redes campus y los precios de sus componentes. Indicaremos aquí solamente los aspectos más relevantes al problema estudiado.

Los conmutadores 10 GE son aún muy costosos y se justifica su uso solamente cuando su alta densidad es imprescindible. La capacidad de conmutación para un conmutador de núcleo de N*10 Gb oscila entre valores de 512 Gbps y 1,68 Tbps, con capacidades de 384 y 900 puertos de 1 Gb respectivamente.

Otro aspecto observado en las propuestas de los fabricantes en lo referente a las topologías utilizadas para las redes campus es que desaparece en algunos casos la capa de distribución, lo que se justifica porque el tráfico interno es bajo porque predomina el tráfico cliente-servidor. Los conmutadores de gama alta disponen de autenticación, que puede basarse en 802.1X, en dirección MAC o basada en web. En cuanto a seguridad

disponen de chequeo de integridad, y facilidades hardware como la detección a nivel ASIC de paquetes TCP SYN, en cuyo caso puede copiarse el tráfico del puerto en cuestión para ser observado (*mirroring).*

En lo relativo a redundancia en la red, se prefieren enlaces duplicados entre Núcleo y Acceso a tener redundancia en el *switch fabric* de los conmutadores propiamente dichos, dado que su fiabilidad (MTBF) es de decenas de años. Ejemplos de las capacidades actuales [Wit05] de conmutadores de gama alta son:

Conmutadores de Núcleo:
-Alcatel Omniswitch 8800. Capacidad total 512 Gbps, hasta 384 puertos full duplex 1 GE. El precio de un enlace 10 GE 15.000 $ en el modelo 8800 más 10.000 $ en el modelo 6.800 suma un total de 25.000 $.
-Extreme Networks: Core Blackdiamond 10808. Capacidad 1,6 Tbps. Hasta 480 * 1 GE ó 48*10 GE.
-Foundry Networks: BigIron MG8. Capacidad 1,28 Tbps. Hasta 320 puertos 1 GE o hasta 32 puertos10 GE.

Conmutadores de Acceso:
-Alcatel Omniswitch 6800: 160 Gbps capacidad
-Extreme: Summit 300 Access: 24 FE + 2*1 GE. Precio aprox.: 6.000 $.

Es de destacar el protocolo de protección EAPS, soportado solamente por un gran fabricante (Extreme), que reconfigura en 50 mseg. según RFC3619. Esto permite evitar cambios de topología cuando un enlace entre nodos cae, si existen varios enlaces agregados entre los nodos mediante protocolo IEEE 802.3ad.

2.9.3 Criterios de costes.

El diseño de redes campus utiliza criterios de coste económico para la toma de decisiones y de métricas para establecer los requisitos y la comparación de prestaciones de las redes. La elección del criterio de coste para comparar redes no es sencilla. De hecho la optimización siempre se realiza respecto a un criterio simple o combinado, resultado de elegir una serie de factores a optimizar y unos factores de ponderación y criterios de elección: coste económico del enlace, infraestructura utilizada/no utilizada, longitud media o máxima de la ruta, etc.

El criterio de coste a aplicar en una red no se puede establecer de forma indiscutible y unívoca dada la diversidad de factores que pueden incluirse en el mismo y la forma de valorarlos de la organización propietaria de la red. La forma de cálculo de esta función de coste por enlace en las redes puede ser compleja y tener en cuenta diversos factores, desde ancho de banda a coste económico de la línea utilizada en cada ruta, carga de la línea, etc. Dentro de los costes económicos de los enlaces utilizados podemos distinguir entre los costes de adquisición, que incluyen el coste de adquisición de los equipos, la infraestructura necesaria y la instalación de los mismos (CAPEX), y costes de operación (OPEX), que tienen en cuenta los costes de mantenimiento y soporte del servicio de la red y los posibles costes de alquiler de los enlaces que no fueran propiedad de la organización. La política de cálculo de costes de operación suele ser establecida por la organización propietaria a efectos de reparto de costes entre las diferentes áreas o

departamentos de la organización que permitan costear su mantenimiento y financiar su actualización de forma equitativa, como se justifica en [Vernon]. Estos dos criterios son los que parecen claves, el primero orientado a los costes reales de operación de la red y el segundo a cubrir los costes futuros de renovación de la red. Estos criterios deben ponderarse a fin de combinarlos en uno único al efecto de comparar alternativas de diseño concretas.

Los costes de operación constan principalmente de los costes de personal para configuración y mantenimiento y los costes de alquiler de equipos y líneas de la red campus. Este segundo componente de los costes es normalmente independiente de la solución planteada dado que se trata de unir diferentes áreas de la red campus con una determinada capacidad de transporte. Por ello, en el contexto del problema planteado en esta Tesis, los costes de configuración y mantenimiento son los decisivos. Dado que los requisitos de las redes campus propuestas incluyen la mínima necesidad de configuración los costes de operación serán sensiblemente menores por lo que en el caso que nos ocupa es de esperar que prevalezcan los costes de adquisición (CAPEX). Podrían justificarse económicamente costes superiores de los equipos autoconfigurables respecto a los convencionales por el ahorro anual de gastos de operación de la red.

2.9.4 Modelo FOLS

Se describe aquí, a efectos de referencia actual del equipo utilizado actualmente y sobre todo de sus costes, el cálculo efectuado para la red objetivo mediante el modelo FOLS [FOLS05]. La Sección de Fibra Óptica en LAN (FOLS) de la Telecommunication Industry Association (TIA) publica un modelo de costes para ayuda en el diseño de redes campus, que permite comparar costes de redes de tres configuraciones: jerárquica en estrella, fibra centralizada y FTTE (Fiber To The Enclosure) en alta y baja densidad. En la configuración jerárquica en estrella se utilizan conmutadores de núcleo centralizados y conmutadores de grupo de trabajo de capa 3 en cada planta unidos mediante fibra al núcleo. En la configuración de fibra centralizada estos conmutadores se alojan todos de forma centralizada. En el caso FTTE se sitúan equipos pequeños junto a cada grupo de hasta 8 usuarios en lugar de conmutadores de grupo de trabajo en los repartidores de cada planta. El usuario puede aplicar su propio modelo de costes. En la Figura 19 se muestra la red en estrella jerárquica del modelo y en la Tabla 2 los costes aplicados por defecto, que son indicativos de los precios actuales de los componentes de redes campus. No se contempla concentración en la red por lo que el equipo resultante puede ser excesivo, sobre todo en configuraciones grandes. Los resultados para la red objetivo de 20.000 PCs dan:

Resultados para 20000 PCs	Conm. Core	Conmutadores WG L3	GBICs 1000-X
- Jerárquica en estrella	31	1484	3000
- Fibra centralizada	21	928	2274
- FFTE de alta densidad	24	(*) 3710	3710

(*): miniswitch de 8 puertos.

De las tres configuraciones calculadas por el modelo, la configuración más comparable (excepto el importante punto de que no supone concentración), es la de red jerárquica en estrella.

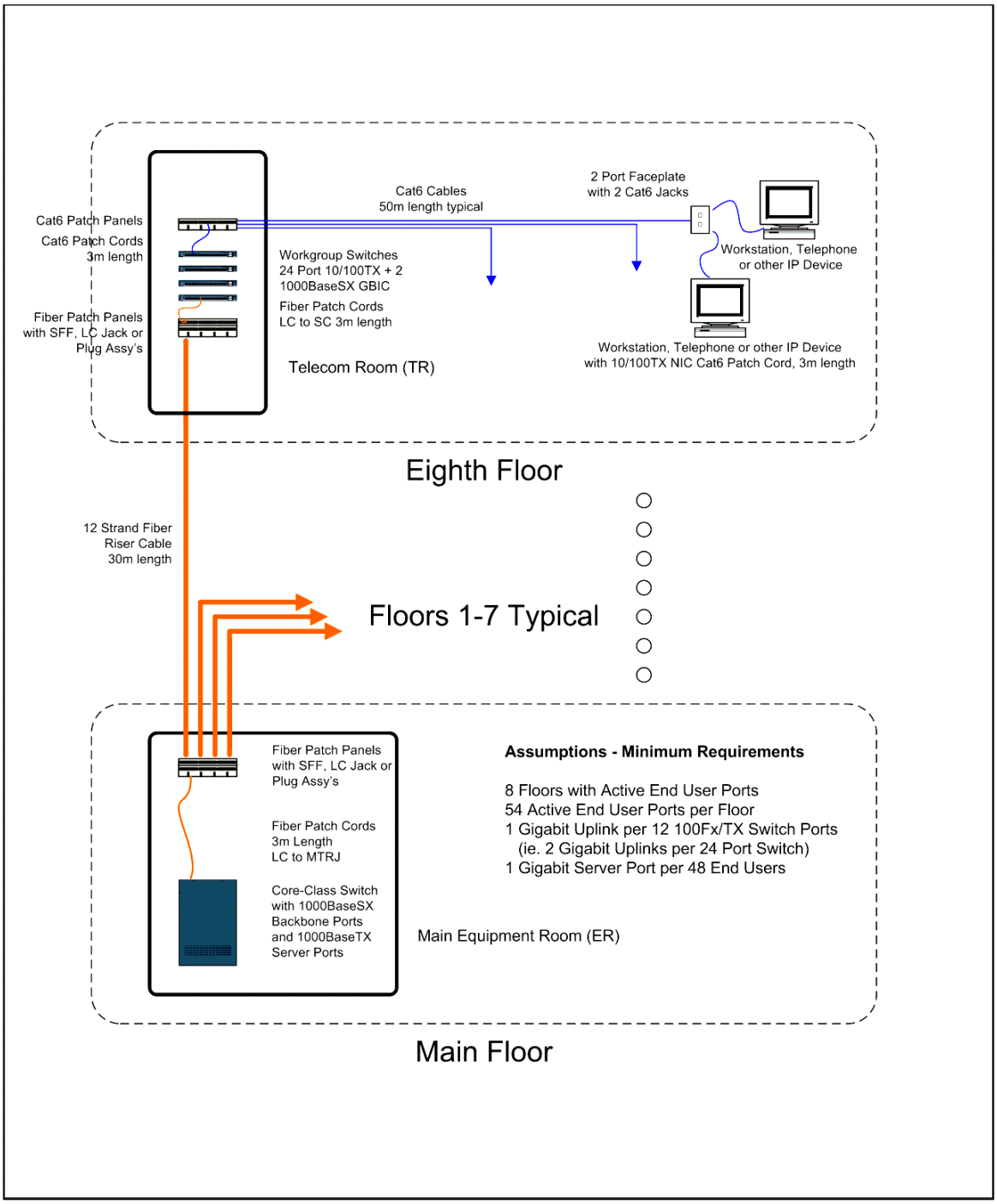


Figura 19. Modelo FOLS de red campus en estrella jerárquica.

Tabla 2. Modelo de costes FOLS 2005

Modelo de costes FOLS 2005
GBIC 1000Base-T……..… 245,73 $ / puerto
GBIC 1000Base-SX……...……...…321,18 $ / puerto
Mini-Switch 8 puertos 10/100Base-T 1*1000Base-SX……..…425,80 $
Switch, Layer 3, 10/100Base-T 24 puertos, 2 * 1000….....1.196,98 $
Switch, Layer 3, 100Base-FX 24 puertos, 2 * 1000 ….…….3.031,65 $
Core Switch hasta 112 puertos. Equipo base…..……..….20.407,09 $
c/16 puertos cobre……...7.713,71 $
c/16 puertos fibra……....8.014,88 $

Los parámetros empleados en el modelo son los siguientes: 54 puertos por planta, 1 puerto servidor cada 24 usuarios. Para una red de 20.000 PCs resultan 31 core switches de tipo N*1 GE y 1.484 conmutadores de acceso con enlaces de agregación 2*1 GbE.

La capacidad total requerida en núcleo resulta:
20.000*100 Mbps*24/2*1000 Mbps= 2*10**6*1,2Mbps =1,67 Tbps.

Con un factor de concentración (sobresuscripción) de 15, se requeriría una capacidad de 111 Gbps. El modelo sirve como referencia de costes de equipos y a efectos de cálculo de equipo necesario si se corrige con el mismo factor de sobresuscripción empleado en la red con la que se compare.

2.9.5 Conectividad de las redes campus.

En este epígrafe se hacen unas consideraciones sobre aspectos de la conectividad que el autor considera pendientes de estudio pero de posible interés en el futuro y se realizan hipótesis sobre los factores que pueden considerarse decisivos. Se trata de un aspecto nuevo de un problema reciente: la ley de distribución del grado de conectividad en las redes de paquetes. Denominamos *grado* de un nodo al número de enlaces del nodo.

Existen múltiples investigaciones sobre la estructura de Internet y la distribución del grado de conectividad en los nodos, pero no equivalentes, que el autor conozca, sobre los grados de conectividad de las redes de ámbito campus. Las simulaciones de redes normalmente [Lui02] asumen unos rangos de variación del grado de los nodos de la red locales simuladas, pero no se aprecia justificación al respecto.

El descubrimiento, relativamente reciente [FFF99], no exento de contradicciones, del cumplimiento de la ley de distribución exponencial inversa (*Power Law*) en Internet en lo que se refiere a la distribución probabilística de los grados de conectividad de los encaminadores ha aumentado los estudios al respecto. En el Apéndice B se explica someramente el contenido de esta ley. Esta aplicabilidad había sido descubierta con anterioridad para otros tipos de redes de tipos muy variados [AB02]. Sin embargo los estudios que conocemos se limitan al nivel de encaminadores, no existiendo (o al menos desconocemos su existencia) algo similar que tenga en cuenta la conectividad dentro de las redes campus a nivel de conmutadores.

Dado que las redes campus se construyen con equipos tipificados y siguiendo las recomendaciones de los fabricantes, es esperable que las topologías de las mismas exhiban una tendencia hacia los modelos de topologías recomendados por los fabricantes, normalmente orientados a diseños jerárquicos de red de tres niveles.

Otro factor de peso esperable en el grado de conectividad es la acumulación y coexistencia de equipos de diversas tecnologías correspondientes a diferentes épocas. En la actualidad lo más frecuente es que coexistan equipos ATM y Ethernet con velocidades de 10/100/1000 Mbps y excepcionalmente, aunque de forma creciente, 10 GE. La tipología de equipos depende mucho de la tecnología, por lo que es razonable suponer como dominante la tipología Ethernet y sus efectos en el grado de conectividad.

Debido a esta coexistencia de velocidades de Ethernet, es frecuente la utilización de conmutadores que agregan la capacidad de enlaces de capacidad inferior, como los conmutadores de grupo de trabajo que se describen más abajo. Los tipos predominantes de conmutadores son actualmente los siguientes:

- Conmutador de grupo de trabajo: Este tipo de conmutadores presenta una configuración de tipo n+2 ó n+1, por ejemplo 24 enlaces Ethernet 100 Mbps más 2 enlaces 1 Gpbs ascendentes que agregan tráfico. Pueden ser apilables, formando varios de ellos un bridge equivalente de k*n puertos siendo k el numero de unidades apiladas, como se muestra en la Figura 20. Los valores de n suelen ser: 8, 16, 24. Si los usuarios están conectados directamente, el grado del conmutador, si incluimos los enlaces sistemas finales, es elevado, pero esto no es relevante a efectos del rendimiento. El grado efectivo en la red es el de interconexión de conmutadores.
- Conmutador n*1: Dispone de n puertos Ethernet de idéntica capacidad (10, 100 Mbps, 1 ó 10 GE).
- Conmutadores rack k*n: Este tipo de conmutadores se configura en rack y equipa un número n variable de tarjetas de k puertos cada una, conectadas a un elemento de conmutación compartido, por lo que resulta en un conmutador de k*n puertos.

La obtención de datos reales de la topología de redes conmutadas actuales presenta dos obstáculos principales: Por un lado, la transparencia de los conmutadores, que los hace invisibles a comandos como traceroute/tracert enviados desde los sistemas finales y por otro lado la confidencialidad de la topología detallada de las redes campus por razones y normativa de seguridad de las organizaciones. Los departamentos de sistemas de información obtienen la topología de sus redes mediante programas de inventariado de equipos a través de la interfaz de gestión de los bridges, pero las razones de seguridad prohiben la difusión de la topología detallada de la red, por lo que una investigación extensa de este tema requerirá la obtención de múltiples autorizaciones administrativas de los respectivos responsables de red.

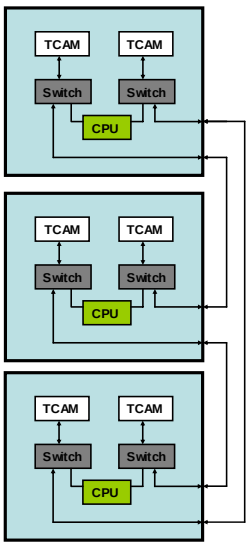


Figura 20. Conmutadores apilables.

Nos parece relevante el estudio de la distribución de la conectividad interna en los dominios conmutados en las redes campus para su aplicación en el diseño, simulaciones y adecuada evaluación de protocolos y dispositivos que operen en capa dos.

2.10 Equilibrado de carga. Redes Valiant.

Las propuestas descritas hasta aquí llegan en algunos casos, como en los protocolos de árboles múltiples de expansión, hasta distribuir la carga, pero sin realizar un equilibrado activo de la misma. La capacidad de reparto de carga para evitar la congestión es una de las ventajas de los encaminadores frente a los conmutadores, mencionada al compararlos en la Definición del Problema. Cara a la posible implementación de mecanismos de este tipo en capa dos, se considera relevante mencionar las redes Valiant [VB81] a efectos de trabajo futuro como posible línea de evolución de los protocolos de troncal propuestos en esta Tesis.

Las redes Valiant fueron originalmente concebidas para la interconexión entre procesadores de altas prestaciones, pero han sido recientemente propuestas para diseño de troncales [ZM04], y para diseño de encaminadores de alto rendimiento [KES03] [CHA01]. Consisten en redes malladas de conectividad lógica total entre todos los nodos. El mecanismo de equilibrado de carga consiste en repartir el tráfico entre cada pareja de nodos de forma equilibrada entre todas las rutas de dos saltos existentes entre ambos nodos. En la Figura 21 se muestra el principio de funcionamiento. Supongamos una red troncal de N nodos con conectividad total y con un tráfico de entrada agregado de su red de acceso de r por cada nodo. El reparto consiste en repartir el tráfico r que entra en un nodo uniformemente entre los enlaces del nodo, independientemente del nodo destino, de forma que el máximo por enlace debido a ese nodo es r / N. En el segundo salto, todos los paquetes se envían al nodo destino. El tráfico originado por el segundo salto es también como máximo r / N, por lo que basta dimensionar todos los enlaces con una capacidad de $2r / N$ para evitar la congestión siempre que el tráfico por nodo sea inferior

a *r*. El reparto puede hacerse paquete a paquete o por flujos (esto último preferible en el caso de bridges para mantener el orden de las tramas).

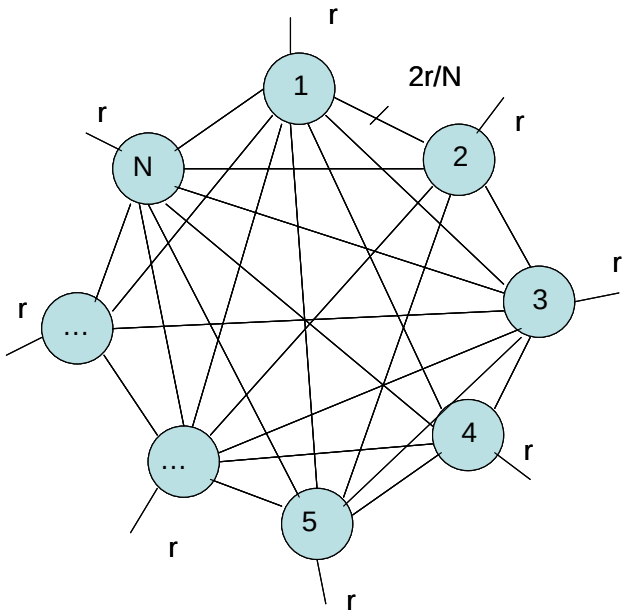


Figura 21. Troncal con equilibrado de carga Valiant.

La característica más destacada es la independencia frente a la matriz de tráfico. Esta es una importante cualidad, dado que la matriz de tráfico es variable en el tiempo e impredecible. Téngase en cuenta la posibilidad de alteración de la matriz de tráfico en cualquier red campus a medio plazo debida a las aplicaciones P2P como Skype, videoconferencia, o teleformación. Esta impredecibilidad supone una dificultad añadida en el diseño de redes, que conduce normalmente a un sobredimensionamiento importante de las mismas. Las redes Valiant presentan las ventajas de sencillez de arquitectura y la garantía de rendimiento predecible ante fallos o ante cambios en la matriz de distribución de tráfico, requiriendo solamente un sobredimensionamiento reducido de la red. Entre los posibles problemas a resolver en su aplicación en capa dos es la alteración del orden de las tramas, ya que al existir caminos alternativos, algunas tramas emitidas con posterioridad a otras podrían llegar al bridge destino antes que las emitidas antes, con el consiguiente mal funcionamiento de algunos protocolos de capa dos. Esto normalmente puede resolverse con mecanismos de distribución de tráfico por origen y asignación a una ruta. Otros mecanismos de reparto de carga recientemente propuestos [GOE05] utilizan costes de enlaces variables en función de la carga del enlace, asimismo de gran interés para estudio futuro.

2.11 Conceptos de direccionamiento y encaminamiento.

Se agrupan en esta sección conceptos básicos de direccionamiento y encaminamiento, clásicos y recientes, con posibilidades de ser aplicables con modificaciones para el problema planteado, aunque no directamente vinculados al mismo.

2.11.1 Direccionamiento y encaminamiento jerárquicos

Para que las redes campus escalen hasta los tamaños de 100.000 sistemas finales parece que podría ser necesario establecer cierta jerarquía en dichas redes. Dado que se parte de no utilizar direcciones IP para el encaminamiento en la subred, parece conveniente una revisión del encaminamiento jerárquico y sus tipos y su aplicabilidad en capa dos.

Para establecer direccionamiento jerárquico puede emplearse o bien lo que se denomina una Jerarquía o bien un Árbol de Expansión. La diferencia principal entre una Jerarquía y un Árbol de Expansión consiste en que la Jerarquía contiene solamente a los nodos participantes de la misma, los cuales no precisan ser contiguos y un Árbol de Expansión contiene y conecta a todos los miembros de forma contigua entre todos ellos. Una Jerarquía utiliza algunos nodos seleccionados, con dependencia entre ellos para establecer el direccionamiento de todos.

Muchos protocolos actuales emplean jerarquización, protocolos como IP, DNS y otros. En el caso de IP la jerarquía comprende inicialmente los dos niveles en los que se divide la dirección IP: Identidad de red e Identidad de host o sistema. Con la introducción de CIDR y las subredes esta separación y la longitud de los prefijos de red se flexibilizan.

El encaminamiento jerárquico requiere el establecimiento de una jerarquía que en el caso de DNS se realiza manualmente configurando cada servidor de un nivel con una lista de posibles servidores padre por orden de preferencia. El establecimiento manual es muy laborioso y es importante dentro del problema planteado, el realizarlo de forma automática. El establecimiento automático de jerarquías puede realizarse de arriba-abajo o de abajo-arriba. El problema de establecer una jerarquía está muy relacionado con el de agrupamiento (*clustering*), correspondiendo el establecimiento de jerarquía arriba-abajo con el *agrupamiento por división* y el establecimiento de jerarquía abajo-arriba con el *agrupamiento agregativo* [TV98]. Los mecanismos más habituales de agrupamiento son los siguientes:

- Construcción de un Árbol de Expansión mínimo y particionado posterior del mismo. Es un caso de agrupamiento por división.
- Asignación de un radio de influencia a cada nodo con difusión dentro de la región. Caso de Landmark [Tsu88]. Este es un caso de agrupamiento agregativo.

2.11.1.1 Jerarquías de área

En una jerarquía de áreas la dirección de nodo corresponde a su posición en la jerarquía. Existen distintas aproximaciones [HAG83] para la gestión de estas jerarquías, basándose normalmente en un nodo primario, designado o elegido automáticamente, que supervisa la formación en cada cluster. El proceso se repite sucesivamente entre los nodos primarios elegidos, construyéndose la jerarquía de abajo-arriba.

Es de destacar el trabajo clásico de Kleinrock [KK77] en cuanto a los resultados esperables de jerarquizar las redes por niveles. En este caso el encaminamiento jerárquico utilizado es el encaminamiento jerárquico por área, cuyos principios se muestran en la Figura 22. Este encaminamiento por área utiliza nodos en cada área que enlazan con otras áreas de nivel igual o superior, reduciendo las tablas de encaminamiento de acuerdo con la correspondiente abstracción de la topología que se produce. Cada nodo tiene una entrada en su tabla para dirigirse a sus nodos de igual nivel y a los de cada área superior en cada nivel superior existente.

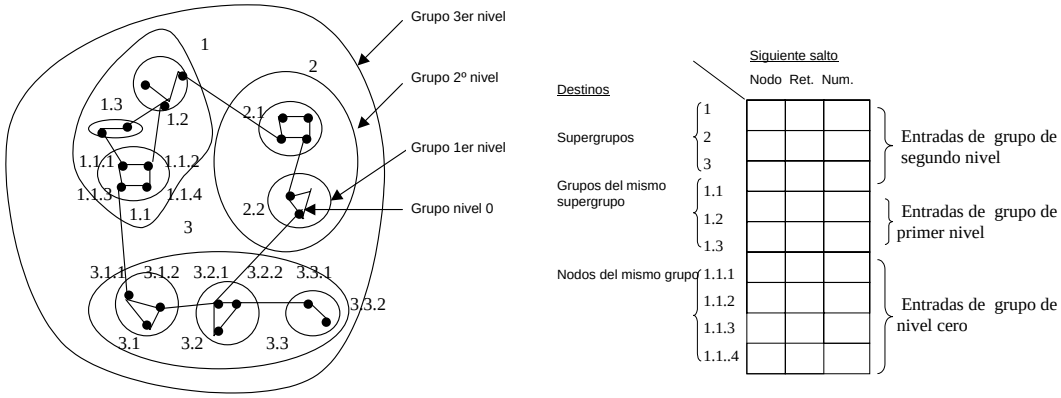


Figura 22. Encaminamiento jerárquico de área (Kleinrock).

El encaminamiento Landmark de Tsuchiya [Tsu88], utilizado en redes *Ad-hoc*, aporta una visión diferente del encaminamiento jerárquico. La dirección de un nodo se determina por la línea descendente desde el nodo raíz y los paquetes se encaminan salto a salto hacia el nodo visible en cada salto que es más cercano a la dirección de destino. Cada nodo difunde anuncios a todos los nodos dentro de un determinado radio, construyéndose de abajo a arriba. El radio varía con el nivel jerárquico y aumenta con el mismo. La elección de nodo padre se realiza limitando el número de nodos por nodo padre a un valor máximo. Esta característica lo hace adecuado para redes radio, en las que la comunicación con el nodo padre comparte el medio con todos los nodos cercanos.

Igualmente son relevantes como protocolos jerárquicos Decnet, IS-IS y OSPF, analizados en [Per00], donde se aborda la problemática del encaminamiento dentro del área, problema que también es parte del encaminamiento jerárquico. IPv4 no dispone de un mecanismo para encaminar dentro de la red Ethernet, por lo que la dirección se resuelve mediante difusión ARP. Por el contrario, protocolos como IS-IS son capaces de encaminar dentro del área inferior con los encaminadores denominados de nivel 1, para lo cual el sistema de direccionamiento (direcciones ISO) incluye la Identidad de Area y la

Identidad de Sistema, empleándose normalmente para Identidad de Sistema los 6 octetos de la dirección MAC, con lo que se incluye explícitamente la dirección de capa dos.

Un protocolo jerárquico específicamente orientado a IP versión 6 es [KS01], en el que se propone la arquitectura denominada GHRA. En ella los niveles jerárquicos son los mismos que los de las direcciones de IPv6: Top Level, Next Level, Site Level. Se emplea el concepto de *Gaterouter* concretado según el nivel jerárquico: Top Level Router, Next Level Router o Site Level Router. Un máximo de 2^{13} routers formarían el nivel superior de Internet con un máximo de hasta 8192 rutas por router de nivel TL. GHRA tiene en cuenta aspectos de encaminamiento IGP y EGP y es compatible con IPv4 e IPv6. El Gaterouter opera como un mini Sistema Autónomo. Estos algoritmos jerárquicos son eficientes en ancho de banda para multicast (se utiliza un router multicast dedicado por cada Gaterouter). El problema del Gaterouter es que el Gaterouter en modo Single no puede procesar todo el trafico y el Multimodo (MMG) es complejo (requiere otro protocolo por encima entre routers del Gaterouter). Es de interés mencionar que esta estructura de direccionamiento en IPv6 ha quedado como histórica. La RFC 3587 de Agosto de 2003 ha dejado obsoleta la anterior RFC 2374 que establecía dichos niveles jerárquicos en el direccionamiento de IP versión 6 por no considerarlos técnicamente adecuados para al despliegue del protocolo, sustituyéndolo por una política de asignación coordinada de las Regional Internet Registries.

En los tamaños de red considerados en esta Tesis, de tamaño significativo pero muy lejos de los tamaños de red de Internet, se puede asumir que la jerarquización es conveniente en las redes campus consideradas, pero de forma menos determinante que en las redes metropolitanas (MAN) o de área amplia (WAN). Aunque cierta jerarquización es necesaria para posibilitar la escalabilidad (la mencionada explosión de direcciones MAC es un ejemplo), una jerarquización simple en dos niveles puede bastar para cumplir los requisitos, pero esta jerarquización debe realizarse a nivel inferior a IP, dado el planteamiento del problema. Para realizar el encaminamiento en capa dos se precisan protocolos de comunicación en capa dos que revelen la conectividad de los conmutadores entre ellos y a los sistemas finales. Si deben ser jerárquicos, deberá usarse un procedimiento de asignación de direcciones en capa dos.

2.11.2 Autoconfiguración y asignación de direcciones.

Los protocolos jerárquicos como IP utilizan un sistema de direcciones jerárquico para realizar la abstracción de la topología y simplificar el encaminamiento. Si la asignación de direcciones es manual, resulta muy compleja y costosa. Ante la necesidad de asignación de direcciones IP en las redes campus se han planteado diferentes soluciones para automatizar el proceso.

2.11.2.1 DHCP y otros asignación de direcciones IP

El protocolo más utilizado actualmente para asignación de direcciones IP a sistemas finales es DHCP [Dro97]. Derivado de Bootstrap Protocol (BOOTP), permite la asignación automática de direcciones IP reutilizables y la configuración de otros parámetros. Para IPv4 se encuentra en estado de Draft Standard, documentado por RFC2131 y RFC2132 y para IPv6 como Proposed Standard documentado por RFC3315. El grupo de trabajo del IETF DHC WG actualmente revisa estos estándares y tiene como primer objetivo resolver la seguridad en DHCP, ante la no implementación ni despliegue

de la propuesta oficial RFC3118). A pesar de facilitar la automatización en la asignación de direcciones y la gestión de rangos de direcciones IP, se precisa asignación manual en el caso de direcciones fijas.

2.11.2.2 Autoconfiguración

El problema de autoconfiguración planteado es específico de capa dos, dado que se asume como punto de partida un segmento único IP en la red y debe respetarse la independencia de capas, lo que excluye la interacción entre el direccionamiento y encaminamiento en capas dos y tres (IP), salvo para optimizaciones opcionales. Existen escasas propuestas de autoconfiguración en capa 2 aparte de las descritas, como Autonet. Dado que existe un cierto número de propuestas en capa 3, se revisa el estado del arte para capa 3 para explorar posibles procedimientos trasladables a capa 2.

Existen algoritmos de asignación de direcciones [CZM00] que descubren la topología a nivel de red explorándola en anchura o en profundidad (*Breadth First Search (BFS)* o *Depth First Search* (*DFS*) y asignan direcciones dentro de la subred IP. Presentan el inconveniente de que la utilización del rango de direcciones IP no es óptima. En [MM00] se propone una combinación de dos protocolos para obtener la autoconfiguración dentro de un dominio. El protocolo de configuración (DRCP) es una ampliación de DHCP y el de asignación de direcciones es un nuevo protocolo (DAAP). Mediante el protocolo DAAP se obtienen los *pools* de direcciones que utiliza DRCP. El proceso de autoconfiguración opera en cascada arrancando de un solo un dispositivo inicial dotado de direcciones IP para asignar.

En el entorno del problema descrito no se utilizan direcciones IP para el encaminamiento sino direcciones de capa 2 (direcciones MAC u otro tipo de direcciones). El uso de otro tipo de direcciones distinta a las MAC presenta serias dificultades de compatibilidad. En cuanto a las direcciones MAC, éstan son únicas, planas y preasignadas (salvo uso de direcciones MAC privadas), por lo que no procede la asignación de direcciones, pero por ser planas no existe posibilidad de agregación. En este caso una alternativa razonable (salvo problemas de escalabilidad) para la autoconfiguración consiste en utilizar un protocolo de estado de enlaces que difunda las direcciones MAC conectadas para encaminar por direcciones MAC. Es el caso de Rbridges. Otras alternativas son las que autoconfiguran construyendo la jerarquía mediante elección de líder desde abajo, como Landmark [Tsu88]. Las existentes en redes ad-hoc, en particular las de tipo jerárquico, son de interés pero de complejidad excesiva para las redes campus. La alternativa de emplear direcciones MAC privadas asignadas topológicamente para facilitar el encaminamiento, presenta posibilidades de implementar un sistema de direcciones MAC jerárquico dentro de la red campus, con el insuperable inconveniente de forzar el uso de direcciones MAC privadas en todos los dispositivos del campus, la restricción de la longitud fija de seis octetos y subvirtiendo el principio de las direcciones MAC como identificadores únicos a nivel mundial.

2.11.2.3 Descubrimiento de topología.

Los algoritmos de descubrimiento de topología se orientan generalmente a facilitar la gestión de red, controlando la topología actual, aspecto esencial para el mantenimiento de las redes debido a los cambios frecuentes del cableado. En muchos casos se concentran en los dispositivos de capa 3 (routers) e ignoran la conectividad de capa 2, con lo que, dada la difusión de los conmutadores, cada vez es mayor la parte de red ignorada. En otros casos son aplicaciones propietarias para determinados dispositivos. [LLC98]

descubre la conectividad IP mediante acceso a las MIB de los encaminadores, ICMP, el programa traceroute y DNS. [BGM+00] describe una implementación basada en SNMP para descubrir la conectividad de nivel 2 y 3 utilizando los datos obtenibles de las MIB de conmutadores y encaminadores mediante SNMP, además de ICMP. Mediante comparación de las MAC aprendidas en las tablas de reenvío de los bridges y mecanismos adicionales se deduce la topología de la red. [LOG01] es un desarrollo ulterior sobre ideas comunes. Los algoritmos de descubrimiento de topología usan SNMP a nivel aplicación, por lo que se descarta su uso por protocolos de capas inferiores para encaminamiento.

2.11.2.4 Descubrimiento a nivel de enlace

El 802.1AB Link Layer Discovery Protocol (LLDP) [LLDP] es un estándar reciente (04/2005) del IEEE para intercambio de información entre dispositivos vecinos conectados a la misma LAN o MAN, relativos a sus características y capacidades. El estándar está orientado principalmente a popular automáticamente las MIB estándar de topología física y de descubrimiento de dispositivos, identificando los dispositivos conectados y las características relevantes a efectos de configuración. Permitirá evitar errores de configuración (p.e. de asignación de VLANs a enlaces). El funcionamiento no satisface las necesidades del problema de direccionamiento y encaminamiento planteados, aunque podría ser un protocolo auxiliar eficaz de otros protocolos de encaminamiento, agrupamiento o asignación de direcciones. El estándar no especifica ningún protocolo de descubrimiento de topología, se centra en el intercambio de información de características entre los equipos vecinos conectados, por lo que es previsible que continúe la situación actual de protocolos propietarios de descubrimiento de topología de los diferentes fabricantes. Aunque parece razonable integrar la funcionalidad de encaminamiento en capa dos con el descubrimiento de topología, la situación actual de los estándares parece ir en dirección opuesta.

2.11.3 Redes Peer-to-Peer.

En esta sección se analiza el estado del arte del direccionamiento y encaminamiento en redes Peer-to-Peer (P2P) y su aplicabilidad al problema. Se ha realizado esta exploración por considerar que los enfoques novedosos de P2P para encaminamiento y direccionamiento pueden inspirar soluciones nuevas en redes campus, pero teniendo en cuenta las importantes restricciones para su aplicación en redes campus.

Una red P2P es una red superpuesta, autoconfigurable, distribuida y descentralizada formada por nodos que actúan como servidores y clientes, compartiendo sus recursos y que se unen a la red y la abandonan de forma dinámica. Las redes Peer-to-Peer (P2P) emplean nuevos conceptos para direccionamiento, búsqueda (*lookup)* y encaminamiento que podrían ser aplicables en redes campus adaptándolos a las diferencias entre ambos tipos de redes. En [Iba03b] se revisaron los principales conceptos y mecanismos de esos tipos utilizados en redes P2P.

Las redes P2P están siendo una fuente de innovación conceptual en redes en los últimos años. Aplicaciones P2P como Gnutella [Gnutella], Freenet [Freenet], Plaxton [PRR97], Groove [Groove], Brocade [Zha+02], y muchas otras utilizan mecanismos nuevos para direccionamiento y encaminamiento. Los conceptos que se aplican son principalmente: Distributed Hash Tables (DHT), Redes Direccionables por Contenido

(Content Addressable Networks (CAN), Distributed Object Location and Routing (DOLR). La visión de las redes P2P consiste en una red miles de millones de ordenadores personales no fiables que presta servicios de forma permanente, escalable, segura y anónima. Se les ha comparado metafóricamente con "la materia oscura del Universo" [PEE01], para visualizar la potencialidad e importancia de toda la capacidad de los ordenadores existentes interconectados y compartiendo sus capacidades computacionales y de almacenamiento. La arquitectura P2P, en la que cada nodo es cliente y servidor simultáneamente en una máquina y aplicación no es nueva, existe en la red telefónica tradicional, en el encaminamiento entre routers IP y en muchas otras redes. Es el predominio del modelo Cliente-Servidor desde que apareció el protocolo http y el éxito de internet en la Worldwide Web el que hace aparecer el modelo P2P como nuevo.

Las redes P2P implementan el encaminamiento en el nivel de aplicación, como red superpuesta (*overlay)*, para poder encaminar a otros nodos sobre la red IP, para lo cual utilizan una identificación de nodos propia y procedimientos de descubrimiento de nodos igualmente propios. Todo ello les permite una gran flexibilidad e independencia de la infraestructura de red.

Existen redes P2P jerárquicas, como es el caso del servicio de nombres de dominios (DNS) en la que los nodos se organizan en niveles y controlan y delegan en otros nodos y existen redes P2P puras como Gnutella en la que todos los nodos son iguales. En otras redes como Freenet, los nodos se especializan y su importancia varía dinámicamente a medida que la red se optimiza, migrando la información cada vez más cerca de sus usuarios para reducir el ancho de banda y la distancia. Es ilustrativa la comparación entre el encaminamiento en redes Peer to Peer y en redes Ad Hoc realizada en [SGF02].

Los conceptos principales utilizados en las redes P2P son los siguientes:

- Red superpuesta autoorganizada desde abajo, en vez de desde arriba.
- Generación automática de identificador de usuario o sistema final. El más frecuente consiste en aplicar una función resumen (hash, por ejemplo SHA-1) de la dirección IP o del nombre del fichero. De esta forma se obtiene un espacio de direcciones de longitud constante y distribución uniforme.
- Se utiliza este identificador para la búsqueda (lookup) de nodos.
- Esta búsqueda es normalmente la aplicación principal de las DHT.
- Proximidad lógica (cercanía entre identificadores) respecto a proximidad física de los usuarios/sistemas. La proximidad lógica permite realizar una búsqueda totalmente distribuida mientras que la proximidad física permite optimizar las prestaciones.
- Reducción de entropía. En algunas redes se optimiza continuamente el rendimiento.

2.11.3.1 Distributed Hash Tables

Las DHT son una clase de sistemas distribuidos orientados a facilitar una infraestructura autoconfigurable y escalable a través de una interfaz de programación simple. Esta infraestructura puede soportar diversos servicios como almacenamiento de datos o encaminamiento y distribución de información. Su nombre deriva de la función de *hashing* utilizada para asignar la propiedad de un dato a un nodo. La DHT facilita a un nodo participante un método de encaminamiento o de búsqueda de datos para determinar la máquina responsable del dato o encaminarse a ella. Los nodos participantes

se reparten la propiedad de un conjunto de claves y pueden encaminar los mensajes al propietario de una determinada clave.

La estructura de nodos en una DHT se basa en una estructura de red superpuesta (*overlay*) con una determinada topología (círculo (CHORD), hipercubo, etc). Cada nodo tiene un identificador lógico que determina su posición en la red superpuesta. Se utiliza un protocolo de asociación al overlay que permite a un nodo nuevo unirse al sistema, contactando a un nodo perteneciente al overlay. El protocolo difunde el nodo a una serie de nodos cercanos y coopera en la construcción de la tabla de encaminamiento del nuevo nodo. La tabla de encaminamiento se usa para determinar el nodo responsable de un dato. Esto se determina asignando una clave al dato (o nodo buscado) en el mismo espacio de identificadores de los nodos y asignando dicho dato al nodo con identificador más próximo (proximidad en los términos definidos por el protocolo, no física). La búsqueda es posible normalmente en O(log n) saltos.

Bajo el término DHT se agrupan sistemas P2P como Chord [Sto+00], Tapestry [ZKJ01], Pastry [DR01] y CAN [Rat+01], que utilizan mecanismos semejantes pero distintos entre sí. La mayoría de ellas utilizan una clave de 160 bitios generada mediante SHA-1 para identificar cada nodo en el espacio de claves. Las parejas de (clave, dato) se almacenan según clave en nodos cuya identidad está cercana a la clave según un determinado criterio de cercanía. La localización de una clave se realiza mediante un algoritmo de encaminamiento basado en la identidad de nodo.

2.11.3.2 Búsqueda de nodo

La búsqueda de nodo es la operación básica para recuperación de datos en los sistemas distribuidos. Es similar a la búsqueda activa de ruta en las redes Ad-Hoc.

Los métodos más utilizados en redes P2P para búsqueda de un nodo son:

- Difusión (broadcast), utilizado por Gnutella y Freenet. Los nodos realizan difusión a todos los nodos conocidos, no se mantiene estructura, pero se consumen muchos recursos y se produce congestión. Equivale al protocolo ARP en IP. Alguna aplicación limita la difusión realizándola primero selectivamente a los nodos más probables. Las respuestas se guardan en cachés en el itinerario de vuelta para ahorrar recursos en posteriores búsquedas.
- Algoritmos deterministas de localización, usado por Pastry y Tapestry. Este segundo método divide la estructura o tabla de búsqueda y se distribuye una parte de la tabla en cada nodo (p.e. búsqueda por sufijo). Estas particiones puede mapearse de forma estática o dinámica en cada nodo. En el mapeado estático cada nodo almacena copia de las particiones cuya clave está en el rango asignado. Si la movilidad es alta, la reorganización será frecuente. Para reducir estas reorganizaciones se puede utilizar mapeado dinámico en el que cada nodo almacena los contenidos utilizados más frecuentemente. Veamos con algo más de detalle la búsqueda *(lookup)* en Tapestry. La búsqueda es determinista, se alcanza seguro el destino, el número de saltos se reduce si se encuentra en el camino algún nodo con el destino en caché. Cada nodo tiene solamente 16 entradas de tabla, se realiza encaminamiento por sufijo como se muestra en la Figura 23, La búsqueda o encaminamiento al nodo 8954 desde el nodo 5230 se realiza en saltos sucesivos, en cada salto se resuelve un dígito hasta alcanzar el destino. Puede igualmente emplearse para localización de datos que para localización de un nodo. Es adecuado para entornos más bien estables. El mecanismo básico de Tapestry

consiste en enviar un mensaje *publicar* al *location root* del objeto. En cada salto, el mensaje almacena en el nodo atravesado información de localización en la forma (*id objeto, servidor*) *(ID(O), ID(S))* que es lo mismo que *clave (hash(IP)*, y localización del nodo (direccion fisica jerárquica). En el nodo los punteros a las réplicas de los objetos se ordenan por cercanía al nodo.

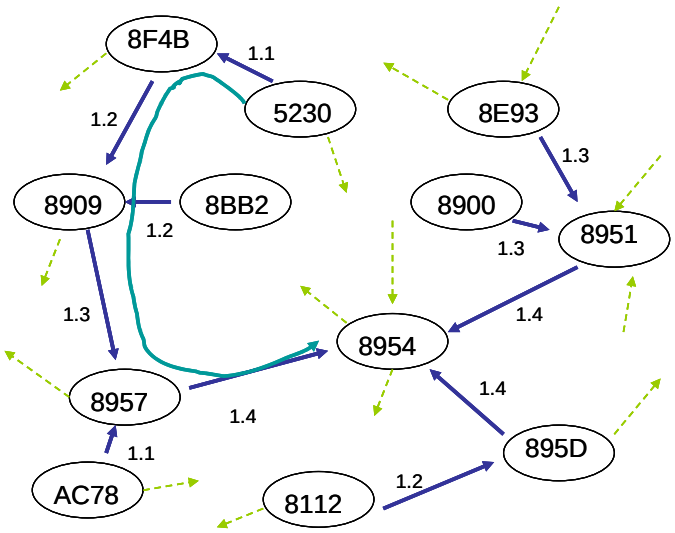


Figura 23. Búsqueda/encaminamiento en Tapestry del nodo 8954.

2.11.3.3 Encaminamiento

Destacan tres métodos de encaminamiento: inundación, búsqueda en círculo (*cycling)* y encaminamiento directo. La búsqueda en círculo es usada por Chord, los nodos están conectados lógicamente en anillo como una lista enlazada cerrada, cada nodo conoce a su predecesor y a su sucesor. Se utilizan primitivas para unirse al anillo y abandonarlo, mediante el hashing se distribuye la carga. El encaminamiento directo se utiliza cuando el nodo conoce la dirección del nodo destino. Se requiere una tabla de encaminamiento en el nodo. El encaminamiento directo no es adecuado en escenarios de alta movilidad por la sobrecarga para mantener las tablas de encaminamiento de los nodos. Si existe jerarquía, los nodos de cabecera de grupo pueden emplear encaminamiento directo y los demás nodos pueden tener movilidad elevada.

Existe el denominado Encaminamiento Jerárquico Dinámico usado en Brocade [Zha+02] que diferencia los nodos creando una segunda red superpuesta de Supernodos, situados en la frontera de los dominios administrativos, para mejorar la eficiencia del encaminamiento. Brocade utiliza Tapestry sobre Tapestry, pero otros modelos jerárquicos similares son posibles como Chord sobre Chord para combinar eficiencia, fiabilidad y escalabilidad.

2.11.3.4 Dimensiones del encaminamiento

Una forma de clasificar los métodos de encaminamiento en las redes P2P consiste en considerar el número de dimensiones en el espacio de encaminamiento. Se distingue así entre encaminamiento lineal, en árbol o multidimensional. Chord es un ejemplo de

encaminamiento lineal (circular) en el que en todos los nodos se utiliza una lista de saltos posibles (saltar a la mitad del círculo, cuarto,etc). Kademlia, Pastry y Tapestry utilizan encaminamiento de tipo árbol. Pastry utiliza una combinación de círculo de identificadores en el que en cada nodo se mantiene un conjunto de nodos terminales, en dos mitades, los inferiores y superiores a *n* respectivamente. El número de búsquedas de la mayoría de los sistemas P2P es de O(*log n*) siendo *n* el número de nodos.

Un sistema que emplea tablas DHT multidimensionales es CAN [Rat+01]. Utiliza un espacio de *d* dimensiones para implementar DHT. El espacio se divide en hiperrectángulos o zonas. Se asocia una zona a cada nodo y el identificador del nodo se forma con los límites de la zona.

2.11.3.5 Localización y encaminamiento Landmark

Landmark es un protocolo para redes móviles relevante por los aspectos de jerarquía y resolución de direcciones. Las redes jerárquicas móviles emplean direcciones que varían al desplazarse el nodo. Esto implica una conversión de la identidad del nodo o dirección original del nodo (*home address*) a la dirección actual del nodo. Para resolverlo, LANMAR [PGH00] [Tsu88] utiliza un servicio de localización de identidad a dirección. Un nodo selecciona su servidor de localización realizando un hash de su propia identidad y utiliza el resultado como la dirección Landmark de su servidor de localización. Cada vez que el nodo cambia su localización, el nodo envía una notificación actualizando su situación al servidor de localización. Para encaminar, cuando un nodo emisor desea enviar un paquete a uno receptor, el enviador calcula el hash de la identidad del receptor n. Pero el hash de la identidad del nodo no coincidirá normalmente con una dirección Landmark utilizable, por lo que la dirección se resuelve nivel a nivel. El nodo envía el paquete de petición hacia el nodo raíz de la jerarquía. Cuando el nodo que reenvía el paquete es lo suficientemente cercano al nodo raíz como para conocer a todos los nodos dependientes inmediatos (*hijos)* del nodo raíz, reenvía el paquete hacia el hijo cuya identidad es más cercana a la de la dirección resumida. El proceso continua a cada nivel hasta que el paquete llega a un jalón (landmark) de nivel 0. Este landmark es el servidor de localización.

2.11.3.6 Peernet

Peernet [EFK03] es una propuesta en elaboración para aplicar conceptos P2P en la capa de red como alternativa a IP. Peernet difiere de otras propuestas P2P en que se utilizan los principios Peer to Peer en la capa de red en vez implementar la red en la capa de aplicación. La capa de red no es IP pero es capaz de interoperar con IP. Peernet utiliza los conceptos de encaminamiento jerárquico de área [KK77] descritos más arriba con un sistema de autodireccionamiento y encaminamiento basado en un árbol binario de direcciones.

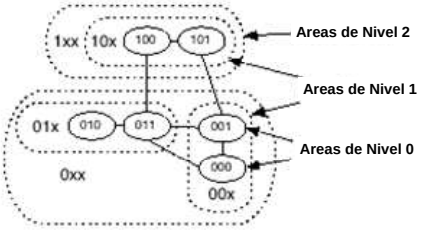


Figura 24. Areas de nivel (agrupaciones) en Peernet.

El sistema de asignación de direcciones en Peernet emplea un mecanismo basado en dividir el espacio de direcciones de cada nodo, como se muestra en la Figura 25. Al unirse un nodo a la red a través de uno perteneciente a la misma se reparten el espacio de direcciones. El nuevo nodo recibe un identificador, el cual es registrado por el nodo en el servicio de búsqueda. En la Figura 25 se muestra el proceso de obtención de direcciones de los nodos A, B, C y D al ir uniéndose sucesivamente en la secuencia indicada para direcciones de tres bits. El nodo A está solo inicialmente y tiene la dirección 000, controla todo el espacio de direcciones de tres bit. Al unirse B a A, divide el espacio de direcciones con B, se le asigna a B la dirección 100 y el nodo A ya no puede asignar direcciones que empiecen por 1. Cuando C se une a la red al conectarse a B, se le asigna la dirección 110 y el nodo B ya no puede asignar direcciones 11x (la 111). Cuando D se une a la red al conectarse a A, se le asigna la dirección 010.

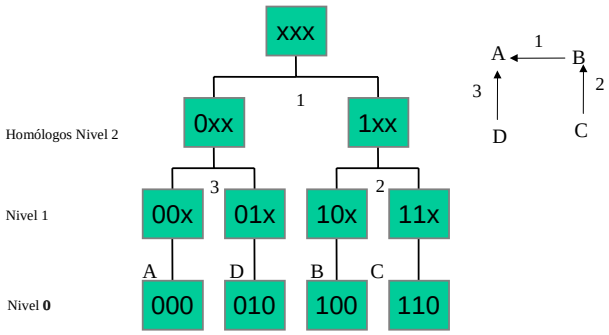


Figura 25. Asignación de direcciones por unión.

2.11.3.7 Métrica XOR

La denominada métrica XOR utilizada en Peernet consiste en calcular la operación XOR de dos identificadores e interpretarla como un entero. Ha sido utilizada en Kademlia [MM02]. Se utiliza un algoritmo de encaminamiento basado en la identidad del nodo que localiza de forma eficiente los servidores cercanos a una determinada clave. El espacio de claves es, de forma parecida a otras tablas DHT, de 160 bit de ancho obtenido mediante hash SHA-1 de datos de longitud mayor, nombre de fichero o dirección IP. La

principal diferencia de Kademlia con Chord es la simetría en la búsqueda, Chord solamente busca hacia delante en el círculo con lo que los nodos ya pasados quedan muy lejanos. Al igual que Peernet, en Kademlia cada nodo es un extremo (*leaf* node) de un árbol binario en el que la posición de cada nodo esta determinada por el prefijo de su identidad de nodo. Para encaminar se requiere que cada nodo disponga de la dirección de un nodo de cada uno de los demás subárboles de la red.

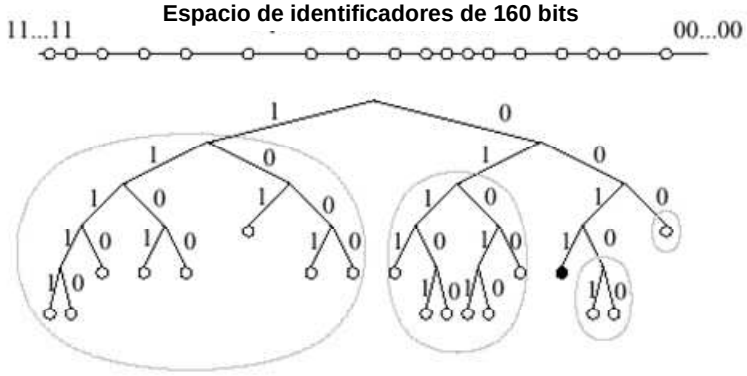


Figura 26. Encaminamiento en Kademlia.

2.11.3.8 El problema de los nodos de cabecera en redes jerárquicas.

En las redes jerárquicas existen normalmente nodos privilegiados que son los de cabecera de grupo, elegidos por los nodos del grupo o por alguno de los superiores, o seleccionados administrativamente para representar todos los nodos del grupo a efectos de encaminamiento así como para asignar direcciones jerárquicas.

Una característica importante de DHT, la de ser totalmente distribuida y descentralizada, lo cual hace posible estructuras sin nodos diferenciados, puede perderse si los nodos de cabecera tienen características diferenciadas de los demás nodos. Peernet resuelve este problema mediante una estructura plana de direcciones de nodos como se muestra en la Figura 25 en la que las direcciones efectivas son las finales del nivel más bajo y las direcciones intermedias de nodos solamente existen virtualmente.

La aplicación en Peernet de conceptos P2P es original y puede ser fértil en el futuro. Presenta sin embargo importantes obstáculos debidos a la incompatibilidad con IP, por la complejidad añadida para interoperar con IP. Es también necesario equilibrar la carga en el árbol y mejorar la conectividad para obtener mayores prestaciones. El factor de agrupamiento de Peernet, al ser binario, crea caminos excesivamente largos en el árbol de búsqueda.

Una diferencia importante de las redes P2P respecto a las redes fijas objeto de nuestro problema es que en las fijas los sistemas finales no encaminan y se busca transparencia respecto a los mismos, mientras que en P2P los sistemas finales o nodos

encaminan, ejecutando un software específico para incorporar la funcionalidad. Esto conlleva modificaciones de los protocolos, por lo que en redes fijas las funciones deberán residir en los bridges y encaminadores en vez de en los sistemas finales, excepcionalmente en algunos servidores. El principal obstáculo para la aplicación de los conceptos P2P de encaminamiento a las redes campus es la lentitud relativa del proceso de búsqueda. Sin embargo, mecanismos como el uso del hash de identificador de nodo para localizar al depositario de una información son válidos y utilizados de forma creciente. En la Tesis se emplea hashing para la obtención del servidor ARP encargado de una dirección IP. Parecen factibles, aunque no se han analizado en detalle, implementaciones Peer to Peer para ARP, pero para evitar la difusión broadcast del paquete ARP en toda la red conmutada es preciso además interceptarlo.

2.12 Optimización de tráfico multicast

En esta sección revisamos el estado del arte de la optimización del tráfico multicast en redes en las capas dos y tres.

2.12.1 Tráfico multicast en capa 2.

El problema de la optimización del tráfico multicast en capa 2 y de su tratamiento por los bridges puede formularse de la forma el siguiente: dado que las direcciones de grupo de destino multicast (GDA) no pueden aparecer en el campo origen de la trama Ethernet, los bridges no pueden aprender estas direcciones multicast. y las direcciones de grupo no aparecen en la tabla de direcciones MAC. Por ello los bridges utilizan inundación para distribuir las tramas multicast. Una opción poco práctica es configurar las direcciones de grupo multicast que se precisen como direcciones estáticas, asociando la dirección de grupo multicast deseada a uno o varios puertos del bridge de forma permanente según corresponda a cada sistema final. Pero esto requiere configuración manual y no se adapta a los cambios.

Existen varias soluciones para este problema en capa 2: La primera consiste en utilizar inspección de mensajes IGMP (*IGMP snooping*) [IGMP]. En esta solución, el bridge intercepta los mensajes del sistema final a su paso por el bridge y actualiza su tabla de destinos MAC añadiendo los grupos multicast pertinentes. Esto requiere un procesado intensivo en los bridges. No todos los bridges soportan esta característica, aunque cada vez está más extendida. Otra solución consiste en que el bridge realice la función de Proxy IGMP, lo que es costoso y requiere procesado. La tercera es la utilizada por protocolos propietarios como CGMP (Cisco Group Management Protocol) [CGMP] que resuelven este problema haciendo que el router multicast se comunique con el bridge de forma que el router interpreta los paquetes IGMP enviados por los sistemas finales e informa al bridge de su contenido, lo que es bastante simple. Finalmente la última y estándar se basa en los protocolos Generic Multicast Registration Protocol (GMRP) y Generic Atribute Registration Protocol (GARP) [802.1D04] [8021Q]. Los routers y sistemas finales emplean dichos protocolos de acuerdo con el estándar IEEE 802.1D para registrarse en los grupos multicast que desean recibir y los bridges procesan estos mensajes si pueden para incluirlos en sus mecanismos de filtrado y reenviar las tramas

con esos destinos por los puertos donde existen suscriptores de dicho grupo multicast. Los bridges que no pueden procesar estos mensajes de registro no los terminan, simplemente los reenvían a los bridges vecinos.

2.12.2 Tratamiento de multicast en capa 3

La definición del problema de optimización del tráfico multicast en capa tres admite una gran variedad de planteamientos, sintetizados en [WH00]. Estos planteamientos abarcan desde el planteamiento clásico del árbol de Steiner [Bla04], cuyo objetivo es minimizar el coste total del árbol de expansión que une a todos los destinatarios del multicast, hasta definiciones que incluyen la optimización de parámetros de Calidad de Servicio (QoS) estableciendo restricciones tales como límites máximos al retardo o a la fluctuación del retardo**.** Estos planteamientos se diferencian principalmente en si las restricciones mencionadas se calculan a nivel de cada enlace o a nivel del árbol. Cuando se establece a nivel de enlace suele ser en lo relativo a la capacidad disponible mínima en el enlace. Las optimizaciones a nivel de árbol se suelen agrupar en dos aspectos: minimizar alguna característica (por ejemplo retardo, o distancia) desde la fuente hasta cada destino, o minimizar las diferencias entre los destinos, como es el caso de las diferencias de retardo entre destinos. Existen asimismo multiples combinaciones de los problemas de optimización simple, combinando optimización de enlaces (múltiples) y/o árboles con restricciones por enlaces y/o por árboles. Por ejemplo un problema de optimizacion de árbol con restricciones de árbol es el de encontrar un árbol de Steiner con retardo máximo acotado. En [WH00] se muestra una visión panorámica de las definiciones de los diversos problemas objeto de investigación en multicast así como de los algoritmos y protocolos utilizados en cada clase de problema, sus ventajas e inconvenientes.

2.12.3 Protocolo de Registro Múltiple (MRP).

Se ha mencionado en párrafos anteriores que el protocolo estándar para gestión de grupos multicast a nivel de enlace es GMRP. GMRP presenta el inconveniente de consumir muchos recursos en su operación para cada sistema final, lo que dificulta su escalabilidad a redes de gran tamaño. Por ello se menciona un estándar en proceso de elaboración, relevante tanto para la funcionalidad multicast como para la de VLANs. El estándar IEEE 802.1ak Multiple Registration Protocol [MRP], se plantea para sustituir al protocolo GARP como protocolo de registro de atributos, el cual es utilizado actualmente para el registro dinámico de sistemas finales a VLANs por el protocolo GVRP [8021.Q] y a grupos multicast por el protocolo GMRP. Esto es debido a que actualmente GARP es el factor limitante en la recuperación rápida en entornos de redes Conmutadas de Proveedor (Provider Bridged Networks) debido al gran número de VLANs registradas en estos entornos. GVRP y GMRP consumen excesivo tiempo y ancho de banda en el registro de la información de VLANs y direcciones MAC multicast. En el borrador de MRP se definen dos aplicaciones MVRP y MMRP para las funciones actuales de GVRP y GMRP. MVRP permitirá además la reconfiguración rápida de la red manteniendo el servicio sin interrupción en las VLANs no afectadas por el fallo.

Capítulo 3

Planteamiento

En este capítulo se plantea el problema a resolver, se realiza valoración del Estado del Arte de los protocolos de capa 2 con vistas a su aplicación en el problema planteado, se expone el proceso de diseño y las decisiones tomadas para definir la arquitectura y los protocolos que se proponen. Se analiza la problemática de autoconfiguración y su relación con las VLAN.

3.1 Definición del problema.

Una definición sumaria del problema que nos ocupa puede ser la siguiente: los modernos conmutadores Ethernet (*bridges*) han hecho posibles nuevos límites para las redes locales, que antes eran de pequeño tamaño y corto alcance. El bajo coste por puerto, alto rendimiento, simplicidad de configuración y la independencia del direccionamiento IP son algunas de sus grandes ventajas frente a los routers. Sin embargo los routers presentan características superiores tales como resistencia a fallos de la red, convergencia, escalabilidad y aprovechamiento de la infraestructura entre otras. Se plantea pues la necesidad de un posible modelo de red capaz de escalar al gran tamaño previsto para redes campus en el futuro a medio plazo (aproximadamente 100.000 sistemas finales), basado en el empleo combinado de dispositivos híbridos que combinen las ventajas de los routers y de los bridges evitando en lo posible sus inconvenientes. Esta red debe aparecer como un segmento único (o subred) para los nodos IP, debe ser lo más compatible posible con los bridges estándar y con los nodos IP actuales tanto con sistemas finales como con routers, sean ambos IP v4 ó IP v6, sin requerir modificaciones

en su funcionamiento respecto a lo establecido en los estándares del IETF rfc1122 [Bra89a] y rfc1123 [Bra89b].

3.1.1 Ventajas e inconvenientes de los bridges

El éxito de los bridges Ethernet descrito en el capitulo primero tiene dos aspectos: el de transmisión y el de conmutación. Por una parte la eficiencia y economía de Ethernet como tecnología de transporte de datos en línea (capa física) y por otra el alto rendimiento y la configuración automática de los bridges Ethernet en el encaminamiento (*bridging)* de los datos entre segmentos . Nos centramos en los aspectos de encaminamiento asumiendo el estado actual de la tecnología Ethernet para transmisión en el nivel físico (de 10 Mbps a 10 Gbps).

Considerando la función de encaminamiento en las redes, las principales ventajas de los Bridges respecto a los routers son las siguientes:

Autoconfiguración. Los Bridges transparentes Ethernet no precisan configuración en su funcionamiento básico. Mediante el mecanismo de aprendizaje hacia atrás (*backward learning)*, aprenden el camino por donde encaminar las tramas, enviándolas por el puerto por donde se recibieron recientemente tramas con esa dirección origen. Aunque algunas funciones de los Bridges actuales requieren configuración detallada, como es el caso de las redes locales virtuales (VLAN), el funcionamiento básico de un bridge no requiere configuración alguna.

Independencia del direccionamiento IP. Como el encaminamiento en la red conmutada no utiliza direcciones IP (ligada al punto de conexión a la red) sino direcciones MAC (ligada al sistema final y permanente), no se requiere modificación alguna cuando un sistema final se mueve de una parte a otra de la red conmutada. Esta es una gran ventaja frente a los routers, que requieren configuración detallada de las rutas y gestión de las direcciones IP de los interfaces. Los bridges soportan de forma automática la portabilidad de los sistemas finales dado que el bridge los detecta automáticamente y actualiza sus tablas al cambiar su punto de conexión en la red campus y emitir tramas desde su nueva conexión.

Por otra parte, los inconvenientes de los bridges son:

Desaprovechamiento de la infraestructura. Por definición, un árbol de expansión consta de un número de enlaces fijo e igual a *N-1* siendo *N* el número de nodos. Siendo *M* el número de enlaces existentes en la red, el número *D* de enlaces deshabilitados por el Protocolo de Árbol de Expansión es:

$$D = M - (N - 1)$$

La utilización de una red es el número de enlaces activos respecto al total de enlaces, viene dada por:

$$U = \frac{N-1}{M}$$

Por lo que cuanto mayor sea el número de enlaces de la red para un número de nodos constante, o dicho de otro modo, cuanto mayor sea el grado medio (*M/N*) de los nodos de la red (número medio de enlaces por nodo), menor será el porcentaje de utilización de la red. En la Figura 27 se muestra un ejemplo de la topología activa resultante del protocolo de árbol de expansión.

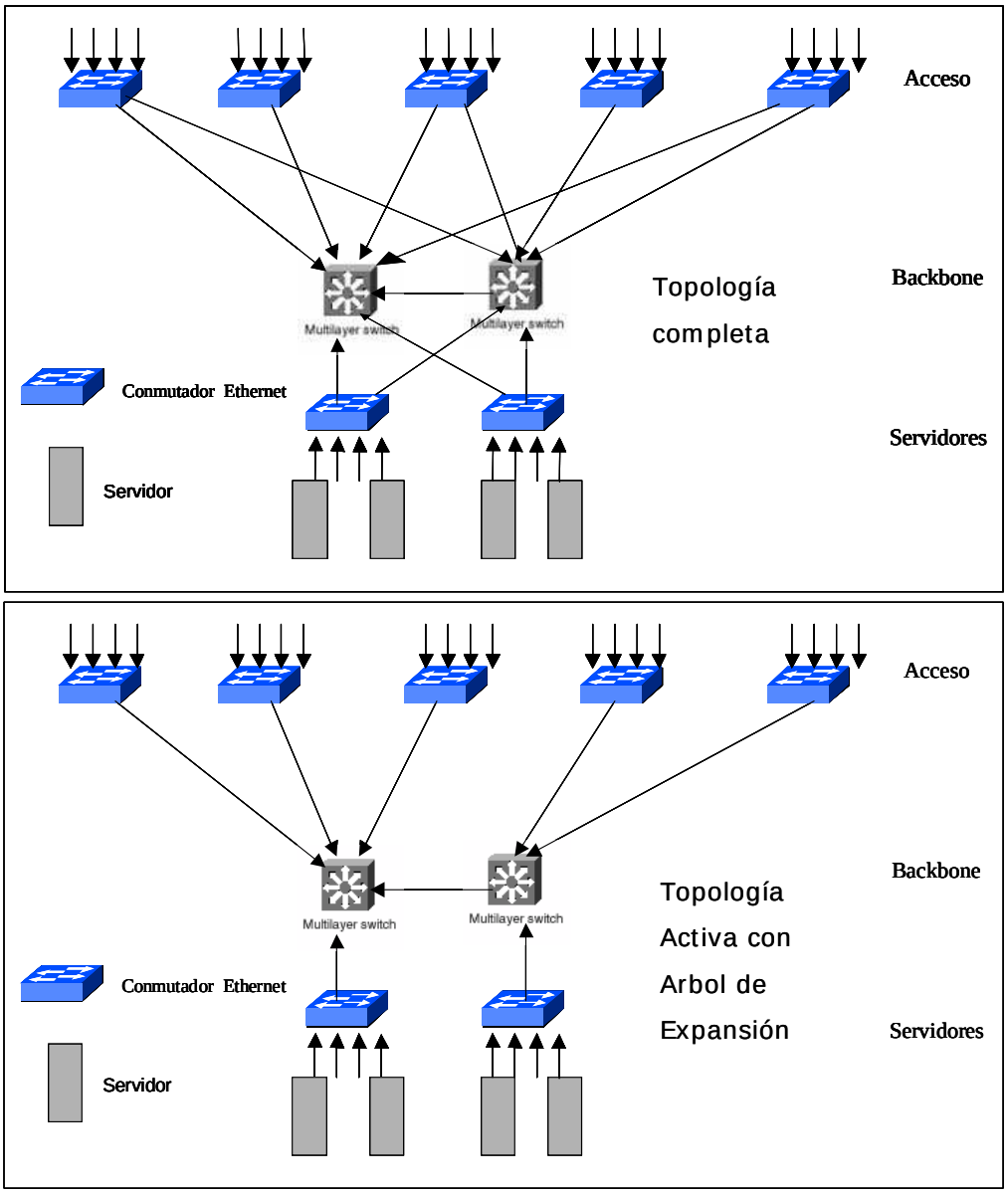


Figura 27. Topologías completa y activa con Árbol de Expansión.

Tiempo de convergencia. Los protocolos del Árbol de Expansión (STP) convergen en 45-60 segundos, frente a los 5-10 segundos de los protocolos de encaminamiento estándar como OSPF. La razón de esta diferencia es que los bridges temporizan antes de habilitar los puertos para reenviar tramas con el fin de prevenir un colapso que podría crearse al habilitar puertos si existen aún bucles transitorios. (Esto se evita en los protocolos de encaminamiento con el campo TTL, del que carecen las tramas Ethernet, por no poder ser alteradas en los bridges). Dado que en muchos casos al bridge no le es posible detectar la finalización del algoritmo es frecuente en los algoritmos de capa dos (como STP) esperar la convergencia del algoritmo mediante temporizadores que esperan un tiempo superior al

caso peor de convergencia del algoritmo, lo que resulta excesivamente lento para los requisitos de las redes para recuperación rápida ante fallos de un elemento. Esta situación cambia con el protocolo de Árbol de Expansión Rápido (RSTP), el cual mplea otros mecanismos basados en decisión local entre bridges vecinos para la habilitación inmediata de puertos alternativos.

Caminos no mínimos. Como consecuencia de utilizarse un árbol único de difusión para el encaminamiento de las tramas por la red, los caminos entre los nodos no son mínimos, siendo solamente mínimos los caminos entre cada nodo y el bridge raíz, y los caminos entre los nodos de cada rama del árbol de expansión, no entre nodos situados en distinta rama del árbol de expansión, lo cuales pueden estar unidos mediante enlaces más directos, pero que son deshabilitados por el protocolo de Árbol de Expansión.

Falta de escalabilidad. Los bridges presentan falta de escalabilidad de las redes formadas por bridges exclusivamente formando un único dominio de difusión. Esta falta de escalabilidad tiene las siguientes causas:

-Proliferación de direcciones MAC. A medida que la red contiene más sistemas finales, los bridges tienen que aprender en las cachés todas las direcciones MAC de los sistemas activos, con lo que éstas se saturan, produciéndose inundaciones innecesarias de las tramas cuyo destino no figura en la caché del bridge.

-Inundación de tráfico de difusión. El tráfico de difusión, como en el caso de los mensajes ARP para resolución de direcciones, se distribuye por todo el dominio de difusión, a través del árbol de expansión hasta todos los sistemas finales. Esto supone una cierta capacidad de la red consumida con esta sobrecarga. Y supone también que todos los sistemas finales deben procesar estas tramas, consumiendo inútilmente capacidad de proceso del sistema final, porque solamente una mínima parte de estas tramas va realmente destinada al sistema final.

-Efectividad del Árbol de Expansión El tamaño de una red de bridges está limitado porque el rendimiento del árbol de expansión disminuye al aumentar su tamaño: los caminos son proporcionalmente más largos y los temporizadores que determinan el tiempo de convergencia aumentarían si se sobrepasan las recomendaciones del estándar en cuanto a dimensiones de la red.

3.1.2 Ventajas e inconvenientes de los routers

Se examinan ahora las ventajas de los routers respecto a los bridges vistas desde el lado de los routers:

Aprovechamiento de la infraestructura. Los routers poseen ventajas innegables frente a los bridges. Estas ventajas son las del encaminamiento en capa tres respecto a los protocolos de capa 2, tales como el aprovechamiento de toda la infraestructura sin necesidad de deshabilitar enlaces. Los protocolos de encaminamiento utilizan toda la infraestructura mientras que los protocolos de capa 2 como los protocolos de Árbol de Expansión (STP y RSTP) deshabilitan enlaces para evitar la formación de bucles en la red, creando un árbol de expansión (*spanning tree)* para la difusión de las tramas, sobre el

que opera el aprendizaje de direcciones MAC por los puertos de los bridges transparentes Ethernet *(backward learning)*.

Prevención de bucles. Otra ventaja de los routers frente a los bridges es la prevención de bucles mediante el uso del campo TTL en los paquetes IP. Cuando se producen bucles en una red conmutada ocasionan tormentas de tramas de difícil diagnóstico y compleja recuperación. Esto se debe a que el bridge, cuando no recibe mensajes del protocolo de árbol de expansión durante un tiempo (p. e. por sobrecarga de tráfico), vuelve a habilitar los puertos para intentar recibirlos. En los bridges los bucles transitorios pueden producir resultados catastróficos por la ausencia del campo TTL. Los routers soportan bien los bucles transitorios (inevitables en los algoritmos distribuidos) no solamente porque tienen un campo TTL, sino porque solamente encaminan en una dirección.

Difusión. El encaminamiento evita la difusión (*broadcast*), que es el punto débil de los bridges. Cuando un bridge no tiene en su caché la dirección MAC destino de una trama recibida, replica la trama en todos los puertos del bridge excepto el puerto por donde se recibió la trama. Esta *inundación* reduce la seguridad de la red al poder ser observadas tramas en todo el dominio conmutado. Aunque las redes locales virtuales (VLAN) crean dominios de difusión separados, son solamente una solución parcial al problema del broadcast.

Tráfico multicast. Los routers procesan por defecto de forma más eficiente que los bridges el tráfico multicast, dado que el comportamiento por defecto de los bridges consiste en difundir indiscriminadamente (*broadcast*) el tráfico multicast recibido. Existen mecanismos para evitar la inundación como la inspección IGMP (*IGMP snooping*) en capa 2.

Rutas óptimas. Los routers permiten la optimización de las rutas o caminos en la red, ahorrando saltos innecesarios respecto a los caminos establecidos por el Árbol de Expansión.

Reparto de carga. Los protocolos de encaminamiento permiten implementar en capa 3 un reparto de carga de forma sencilla, flexible y potente, mientras en capa 2 son de implementación más compleja. La elección de caminos puede hacerse con criterios flexibles utilizando políticas, rutas estáticas, distancias administrativas, etc.

Direccionamiento jerárquico. Las direcciones IP permiten un direccionamiento jerarquizado, que permite agregar rutas. Las direcciones de capa dos (direcciones MAC), son planas y no indican posición topológica.

Compartimentación. Los routers permiten una compartimentación segura de la red campus en segmentos conmutados independientes. De esta forma un fallo en una LAN por mal funcionamiento o error de configuración no afecta a toda la red sino solamente a la red en fallo.

Por el contrario los principales inconvenientes de los routers son:

Necesidad de configuración y administración IP . Los routers precisan configuración de rutas y administración de direcciones IP. Además, si hay cambios de topología en la red o actualizaciones de equipos o enlaces esta configuración debe ser modificada.

Coste superior. Los costes por puerto de un router convencional son alrededor de diez veces superiores a los de un bridge. Sin embargo, los denominados conmutadores de capa tres o conmutadores multicapa (*multilayer switches, switch-routers)* [Anr98] han complicado y alterado el panorama. Existen diferentes tipos de conmutadores multicapa switches, algunos de los cuales han reducido drásticamente la diferencia de precio con los routers, permitiendo implementar subredes IP y encaminamiento IP (pero normalmente sin encaminamiento WAN) dentro de las redes campus para compartimentarlas. Estos switch-routers tienen precios más cercanos a los de los switches convencionales, aunque con restricciones funcionales respecto a los routers estándar. Por ello el peso de este inconveniente no se considera determinante en este contexto de aplicación.

3.1.3 Conceptos de *bridging* y de *routing*

Existe confusión sobre el alcance de los términos *routing* y *bridging.* La evolución de las redes en los últimos treinta años no ayuda a delimitar claramente lo que puede entenderse por *bridging* respecto a lo que puede entenderse por encaminamiento (*routing)*. Radia Perlman [Per05] expone lo que considera encaminamiento respecto a lo que considera bridging de la siguiente forma: el encaminamiento es transporte (*relay)* en capa 3 (capa de red), mientras que el *bridging* es una función idéntica pero realizada en capa dos (capa de enlace). La diferencia es que en la capa 2 se trata de comunicación entre dispositivos vecinos y en la capa 3 es entre diversos sistemas finales. La confusión actual en cuanto a qué funciones corresponden a la capa 2 y cuáles a la capa 3 es resultado de la evolución de las tecnologías y estándares de red que, desde el bridge original que enlazaba solamente dos segmentos de LAN, llevaron a su reinvención en forma de "switch" (conmutador o bridge multipuerto), acompañado por el protocolo de árbol de expansión, lo que permitió operar sin necesidad de configuración (*plug and play*) y de forma compatible con distintos protocolos de encaminamiento de capa 3 (IP, IPX, Decnet). De esta forma el uso de Ethernet se "extralimitó" desde la tecnología de enlace que es (sin campo TTL ni otras funcionalidades) a algo cercano a una tecnología de red. Dentro de esta "extralimitación" de Ethernet se inscribe la utilización de Ethernet en entornos MAN y WAN.

En términos estrictos, las principales diferencias que entre capas dos y tres se pueden considerar son: en cuanto al direccionamiento, que en capa 3 las direcciones son jerárquicas (IP) y tienen significado topológico, estando asociadas al punto de conexión a la red, mientras que en capa 2 las direcciones son planas e identifican al equipo conectado, asegurándose su unicidad (EUI-48) y permanencia en el tiempo; en cuanto al formato de trama, en capa 3 las direcciones incluyen un campo TTL y otros campos para fragmentación de paquetes, opciones de encaminamiento, etc. En capa 2, el formato de trama no incluye campo TTL, entre otras razones porque las tramas no deben ser modificadas por los bridges.

3.2 Problemas a resolver

Consecuentemente con el planteamiento del problema, los problemas principales a resolver, vistos desde el lado de los bridges, ya se han descrito y coinciden con los

inconvenientes de los bridges: Tiempo de convergencia, caminos no mínimos, infrautilización de la infraestructura y falta de escalabilidad. Los tres primeros problemas apuntan a utilizar un encaminamiento alternativo al árbol de expansión. Este encaminamiento podría optimizar o mejorar todos ellos. La falta de escalabilidad, sin embargo, es consecuencia de varios problemas como se ha visto, aparece cuando se trata de aumentar el tamaño de los dominios conmutados y se muestra en los problemas que describimos a continuación.

3.2.1 Problema de difusión ARP.

En las redes campus actuales, el tráfico que utiliza mecanismos de difusión consume un ancho de banda mensurable y tolerable. En la medida en que se incrementa el tamaño de los dominios conmutados como es el caso que nos ocupa, este tráfico se acumula en toda la red y se incrementa linealmente con el número de dispositivos. Por tanto uno de los requisitos en las redes campus es evitar la difusión (*broadcast*) excesiva en la red, porque puede consumir gran capacidad de proceso de los equipos, ancho de banda no despreciable y reducir la seguridad. El protocolo ARP utiliza el mecanismo de difusión (*broadcast)* para interrogar a todos los dispositivos conectados a la red local a fin de obtener la dirección de nivel dos correspondiente a una dirección IP. Estos paquetes deben ser procesados por todos los sistemas finales. En las redes conmutadas muy grandes, esto supone un ancho de banda significativo consumido en toda la red por la difusión y, lo que es comparativamente más costoso, tiempo de procesado en todos los sistemas finales. Esto limita actualmente la escalabilidad de ARP.

La razón de la existencia del protocolo ARP, con los inconvenientes que se acaban de mencionar, estriba en una característica del protocolo IP que consiste en que IP resuelve el encaminamiento hasta los enlaces, hasta la red local, pero no encamina hasta el mismo nodo de destino final. Las direcciones IP son direcciones de enlace, identifican topológicamente un punto de conexión a la red, no identifican a un determinado dispositivo, lo cual hace necesario un encaminamiento adicional hasta el sistema final, que se resuelve difundiendo mediante ARP en toda la subred IP para resolver la dirección del IP del destino y obtener la dirección MAC destino. El protocolo IP en este aspecto no es tan efectivo y coherente como el encaminamiento *hasta el host* de que disponen protocolos como CLNP apoyándose en el protocolo ES-IS [Per00].

3.2.2 Proliferación de direcciones MAC *(MAC explosion problem*).

Cuando el dominio de difusión en una red de bridges es muy grande, los bridges deben aprender todas las direcciones de los sistemas finales activos a fin de evitar inundaciones innecesarias en la extensa red. Si el número de sistemas finales es muy elevado supera el tamaño de la memoria caché del bridge. Las tablas de los bridges se saturan, lo que obligará a la inundación de las tramas recibidas en el bridge cuyo destino no figure en la caché, aumentando la sobrecarga de tráfico. Esta proliferación podría resolverse por mecanismos de agregación de las direcciones aprendidas si éstas tuvieran estructura jerárquica, pero no es el caso de las direcciones MAC Ethernet.

3.2.3 Problema del tráfico multicast

Un aspecto importante en las redes campus es la eficiencia de las mismas para la distribución de tráfico de difusión múltiple (*multicast).* Tanto los servicios que en número creciente utilizan multicast para distribución de contenidos multimedia como los protocolos que utilizan direcciones multicast requieren un uso eficiente de la infraestructura. Al ser la red de bridges considerada de tamaño muy superior a las actuales, puede incurrir en ineficiencias si la red, por trabajar como una subred IP única, inunda de forma indiscriminada en capa 2 para las tramas cuyas direcciones MAC destino son direcciones de grupo multicast.

3.2.3.1 Direcciones multicast

Se hace a continuación un breve repaso a la implementación de las direcciones multicast en capa dos y en capa 3. Una definición adecuada del multicast es la de [SM00] "…la capacidad de una red de recibir un mensaje de una aplicación y entregar copias del mensaje a múltiples destinatarios en diferentes lugares". Las direcciones IP versión 4 de destino de grupo multicast son direcciones clase D en el rango desde 224.0.0.0 hasta 239.255.255.255. El mapeado de direcciones multicast IPv4 a Ethernet es tal que solamente se mapean sobre la trama Ethernet como dirección MAC los últimos 23 bits, por lo que los grupos IP multicast con idéntica terminación son recibidos y procesados por la tarjeta de red Ethernet de forma equivalente y pasan al nivel IP del sistema final, donde son filtrados. Para cada dirección de grupo multicast *(Group Destination Address GDA)* existe una MAC asociada. Esta MAC asociada se forma con 01:00:5E:XX:XX:XX, donde XX:XX:XX son los 23 últimos bits de la dirección GDA de grupo IP multicast en hexadecimal. Como ejemplo: la GDA 224.10.10.10 corresponde a la dirección MAC 01:00:5E:0A:0A:0A. Para ilustrar la conversión o correspondencia de dirección de grupos IP multicast a direcciones MAC, los primeros cuatro bits corresponden a los bits de clase D. De los 28 bits restantes de la dirección IP de 32 bit, los cinco siguientes no son utilizables porque no se adquirió en su momento el OUI para las posibles direcciones MAC de ese rango. Por ello las direcciones MAC se forman con los 23 restantes y resulta un solapamiento de 32 grupos multicast posibles sobre cada dirección MAC. Este solapamiento impacta el rendimiento en las tarjetas receptoras, las cuales procesan tráfico de forma innecesaria.

3.2.3.2 Tipos de tráfico multicast

Históricamente, el direccionamiento multicast se ha utilizado principalmente para dos tipos de aplicaciones: para protocolos de red como OSPF o STP en los que se hace llegar a todos los dispositivos de un tipo determinado una información independientemente de su localización pero de forma diferenciada y para anuncios o búsquedas de servicios (p.e.: ARP, DHCP,…). En los anuncios de servicios se trata de reemplazar la siempre costosa información y búsqueda o descubrimiento de servicios por parte de los clientes de una red, que pueden visitarla ocasionalmente lo que hace improbable e inestable el conocimiento de las direcciones de los servicios, por un anuncio periódico de servicios por parte de los servidores que los ofrecen. En IPv6 existen protocolos como Neighbor Discovery que emplean multicast a nivel de segmento IP limitando así la difusión.

En ambos casos el consumo de ancho de banda es relativamente reducido dado que los períodos de envío de la información son relativamente largos, en el caso de STP es de

una BPDU cada 2 segundos por enlace, lo que significa un 0,0003 por ciento para un enlace Fast Ethernet. Los protocolos de encaminamiento y anuncio de servicios operan de forma similar con lo que el ancho de banda consumido por este tipo de tráfico multicast disminuye proporcionalmente a medida que la velocidad del enlace aumenta. Y la evolución de Ethernet desde 1973 sabemos que ha significado aumentos de casi cuatro órdenes de magnitud, especialmente acelerados en los últimos años con las velocidades de 1 y 10 Gigabit por segundo.

Por lo tanto el problema que plantea el tráfico multicast no proviene de este tipo de tráfico sino de las aplicaciones que utilicen multicast de forma intensiva. En este aspecto las aplicaciones de distribución de contenidos como video, intercambio de archivos, etc. tienen un gran potencial de generar tráfico con anchos de banda de varios Megabit o más por segundo por cada origen de tráfico. Las aplicaciones de teleformación, comunicación multimedia corporativa, congresos on-line, periódicos multimedia y similares cada vez son más frecuentes y es previsible su aumento a medida que tanto los proveedores ISP como las redes campus pueden multiplicar su capacidad de forma económica.

Para completar la descripción de la problemática de multicast es interesante destacar que el funcionamiento por defecto en los bridges de inundar por todos los puertos el tráfico multicast tiene implicaciones serias para el rendimiento de la redes, dado que cada trama enviada en multicast consumirá ancho de banda en toda la red, independientemente de la localización del origen y destino del tráfico. En el hipotético caso peor de que todo el tráfico fuera multicast, el enlace más lento de la red limitaría la capacidad máxima de la red.

3.2.3.3 Multicast en redes campus grandes

Para concluir, en el caso de las redes campus que nos ocupan, los aspectos principales a optimizar en el tráfico multicast son los siguientes: limitar la difusión del tráfico multicast de las aplicaciones hasta los sistemas finales usuarios del mismo, utilizando la mínima infraestructura posible para transportarlo, emplear rutas óptimas en este encaminamiento y finalmente distribuir el tráfico en lo posible entre los enlaces para una utilización óptima de la infraestructura previniendo en lo posible la congestión. Se supone que el origen del tráfico multicast puede ser en principio cualquier sistema final. Como criterios de diseño básico, los objetivos manifestados al principio de simplicidad de configuración y economía son preferibles frente al objetivo de optimización absoluta del tráfico multicast, debiendo ponderarse la importancia de la mejora de la utilización de las redes y optimización de caminos con los costes respectivos de las infraestructuras optimizadas, por lo que dicha optimización debe producirse prioritariamente en las infraestructuras más costosas de la red.

3.3 Análisis del espacio de soluciones.

Una vez revisado el estado del arte se procede a su valoración y al análisis del espacio de soluciones para proceder después a la descripción de las decisiones de diseño.

Los objetivos establecidos en la definición del problema consisten básicamente en combinar la autoconfiguración de los bridges con las ventajas de los encaminadores. Se trata pues de mejorar las prestaciones de los bridges en eficacia de utilización de la

infraestructura, caminos óptimos en la red, convergencia rápida y protección ante fallos de configuración. De acuerdo con lo expuesto en el Estado del Arte, los planteamientos para superar la baja utilización de la infraestructura del Protocolo de Árbol de Expansión (STP) pueden agruparse en las siguientes categorías:

- Habilitación de enlaces cruzados entre bridges que ejecutan un protocolo adicional o extensión de STP, como STAR y DLS.
- Uso de árboles de expansión múltiples en vez de árbol único. Es el caso de MSTP.
- Protocolos de encaminamiento sobre direcciones de capa dos: de tipo de estado de enlaces, vectores distancia u otros. Ejemplo: RBridges, LSOM, Autonet.
- Combinaciones de bridges y routers: conmutadores multicapa.

Cada una de estas categorías presenta ventajas e inconvenientes propios. Como principales inconvenientes de los árboles múltiples de expansión hemos visto la gran complejidad de configuración de MSTP y por otro lado la no optimización de los caminos entre bridges. Como ventajas, MSTP permite una utilización mucho mayor de la infraestructura. Los protocolos de árboles de expansión tanto único como múltiple presentan también la ventaja de sus bajos requerimientos por nodo, dado que solo almacenan distancias al/los nodo(s) raíz y no precisan conocer la topología de la red completa. La rapidez de reconfiguración de RSTP y MSTP lo hace aceptable en redes campus, mientras que para redes metropolitanas puede ser cuestionable por no garantizar la reconfiguración en 50 milisegundos característica de los equipos de jerarquía digital síncrona (SDH).

En lo que respecta a los protocolos que emplean enlaces cruzados entre bridges como STAR, presentan la ventaja de compatibilidad con los bridges estándar y una mejora en la longitud de caminos, pero no logran caminos óptimos comparables a los de protocolos de camino mínimo. Su escalabilidad a redes grandes es limitada: en el caso de STAR por el procesado extra de direcciones MAC requerido.

Los protocolos de encaminamiento sobre direcciones MAC del tipo de estado de enlaces presentan las ventajas de relativa rapidez de convergencia y el inconveniente del volumen de procesado requerido por bridge, lo que puede limitar la escalabilidad en redes muy grandes. Los protocolos de encaminamiento con direcciones MAC de tipo vector distancia presentarán menor velocidad de convergencia que los de estado de enlace pero mayor consumo de ancho de banda para comunicarse las tablas de encaminamiento completas. Estos problemas de escalabilidad se evitarían si se jerarquizaran las direcciones. Pero las direcciones MAC son *identificadores* planos, no jerárquicas y por tanto no agregables, aumentando el tamaño de las tablas de encaminamiento de forma lineal con el número de sistemas finales.

Para terminar esta valoración, los conmutadores multicapa son la solución predominante actualmente para algunas funciones específicas en las redes campus, en concreto en la capa de Distribución para realizar la segmentación de la red. En el caso de los conmutadores multicapa basados en *caching*, suelen consistir en implementaciones propietarias y presentar problemas de compatibilidad. Los conmutadores multicapa que actúan como verdaderos routers, denominados a veces *switching routers*, son compatibles con los protocolos RIP y OSPF. En todos los casos requieren configuración y gestión de las direcciones IP de las interfaces y de los segmentos de red involucrados. En síntesis, podemos concluir:

- Si se utiliza encaminamiento basado en direcciones MAC, el tamaño del dominio encaminado debe limitarse mediante compartimentación y también deben buscarse formas de acelerar la convergencia si se requiere inferior a la esperable (Hello time 5 seg.) de los protocolos usados en capa tres, aumentando la frecuencia de los intercambios de mensajes o modificando el protocolo.
- Si se utilizan árboles múltiples de expansión, deben optimizarse los caminos y simplificarse la configuración. El caso peor del tiempo de reconfiguración obtenibles será probablemente similar al de RSTP (1-2 seg.).

3.4 Proceso de diseño.

En esta sección describimos brevemente el proceso de diseño seguido, y una síntesis de las consideraciones realizadas y decisiones tomadas.

Como decisión inicial de diseño, incluida en los requisitos, se asume un segmento único o subred IP en toda la red campus que evite la necesidad de cambio de dirección IP cuando un sistema final se traslada de un punto a otro dentro de la red campus. Para ello es necesaria la eliminación de los routers IP salvo en los bordes de la red campus.

3.4.1 Problema de encaminamiento y autoconfiguración en segmento único IP.

El problema planteado consiste en encaminar en capa dos entre sistemas finales conectados en un segmento único IP. Para realizar el encaminamiento se precisa un sistema de direccionamiento en capa dos. Se exploraron varias posibilidades: encaminar utilizando direcciones MAC, utilizando direcciones IP asignadas automáticamente y utilizando otras direcciones de capa dos.

El encaminamiento basado en direcciones MAC estándar presenta los inconvenientes mencionados de no agregabilidad y de que la dirección MAC es un identificador y no un localizador *(locator)*, no tiene contenido topológico, por lo que no es agregable. En principio aparece como poco adecuado a redes campus de gran tamaño.

Otra alternativa consistiría en asignar automáticamente direcciones IP a los sistemas finales y encaminar con ellas en los bridges. Para utilizarlas de forma agregada en el encaminamiento en los bridges, éstas han de ser asignadas según su conectividad a cada bridge, de forma que cada bridge pueda anunciar un rango de direcciones IP. Para ello se requiere descubrir la topología, para lo cual es preciso normalmente el uso de SNMP e ICMP, protocolos de capas superiores a dos o bien protocolos que elaboren la conectividad de capa dos mediante mensajes entre los bridges, es decir protocolos de encaminamiento de capa dos. Para mantener la independencia entre capas y del protocolo de red, tampoco es aconsejable el empleo de información de IP para encaminar en capa

dos, salvo optimizaciones opcionales. Además deben contemplarse los casos de direcciones IP fijas. Se excluye pues el uso de direcciones IP.

3.4.1.1 Asignación de direcciones de capa dos

Se analiza ahora la utilización de direcciones de capa dos alternativas a las direcciones MAC.

Como principios básicos, recordemos que existen dos esquemas básicos de creación de jerarquía, y por tanto de asignación de direcciones: de arriba abajo y de abajo a arriba. En los sistemas de arriba abajo, es preciso dividir el espacio común de direcciones entre los bridges conectados, con arreglo a su conectividad (número de sistemas finales). En los sistemas de abajo-arriba hay que asegurar la coordinación de las direcciones asignadas por los bridges. Respecto a la asignación de direcciones, dado el aprendizaje de direcciones en los bridges y el requisito de transparencia para los sistemas finales que se traduce en que éstos no manejan las direcciones de capa dos adicionales, bastaría en principio asignar direcciones y encaminar hasta el nivel de bridge designado del sistema final correspondiente.

Para poder agregar rutas en capa dos sería preciso disponer de un *localizador* de capa dos para utilizar en el encaminamiento. El *identificador* es la dirección MAC del sistema final. La clave del éxito de las direcciones IP radica en que se usan como identificador y como localizador. Por ser localizador, permiten la agregación de rutas. Por ser identificador es posible el encaminamiento. Pero en nuestro caso, al ser distintos en capa dos, deberá resolverse la dirección MAC para obtener el localizador de capa dos, el cual indicará el punto de conexión del sistema final a la red, variable si el sistema final cambia de sitio. Si se utilizan direcciones de enlace, esto implica que cada interfaz de red debe recibir una dirección. Y estas direcciones deben ser asignadas automáticamente para que las redes sean autoconfigurables.

Se consideró inicialmente un sistema jerárquico de encaminamiento y autoconfiguración basado en un esquema basado en direcciones MAC anidadas jerárquicamente. Esta aproximación presentaría las ventajas de agregabilidad de rutas y las dificultades de la excesiva longitud de direcciones MAC y la longitud variable de las direcciones. Este esquema de direccionamiento se acerca más a los bridges de encaminamiento en origen que a los bridges transparentes al utilizarse direcciones de bridge en vez de direcciones MAC de puertos de bridge, y hace difícil la compatibilidad con los bridges transparentes. Por todo ello se descarta este esquema de direccionamiento (utilizado por otra parte en algún protocolo de encaminamiento jerárquico de redes Ad Hoc [Iwa+99]).

Durante el proceso de búsqueda de tipos de direcciones jerárquicas de capa dos se contempló la posible utilización del protocolo RSTP para generar las coordenadas topológicas RSTAA, que se describen en el capítulo siguiente. Dada la generalización del estándar Ethernet y la existencia de las correspondientes direcciones estándar de capa dos (MAC) fuertemente implantadas, resulta irrealizable reemplazar las direcciones MAC por direcciones alternativas de capa dos salvo en redes propietarias, por lo que la función de una posible dirección alternativa de capa dos no puede ser exclusiva sino complementaria.

3.4.2 Convergencia del encaminamiento

Un requisito importante de las redes campus es la necesidad de una convergencia rápida del encaminamiento ante alteraciones de la red por caída de nodos o de enlaces.

Un argumento habitual contra RSTP es que el tiempo de convergencia tras reconfiguración es excesivamente largo, lo que es cierto comparado con los equipos SDH, pero no respecto a los algoritmos de encaminamiento de estado de enlaces como Djikstra. Analicemos, siguiendo a [AJY00], la problemática de minimizar la convergencia de estos protocolos de encaminamiento:

- Tiempo mínimo de detección del fallo. Si la detección se realiza a nivel de enlace es muy rápida, pero esto no siempre es posible, por lo que se precisa detectar la pérdida de tres mensajes Hello. La minimización del intervalo entre mensajes Hello de los actuales 5 segundos es un problema de ancho de banda y procesado consumidos. Este componente de tiempo establece un límite inferior para la convergencia.
- Debe comunicarse el fallo a todos los nodos de la red de forma inmediata sin esperar a recalcular la tabla de encaminamiento: tiempo de propagación en la red.
- Deben recalcularse las rutas. Complejidad entre $O(n^2)$ y $O(n*\log(n))$ para Djikstra. Dada la complejidad, si el número de nodos es muy grande debe replantearse el protocolo a utilizar (incremental).

Frente a estos compromisos entre complejidad y rapidez de convergencia, la reconfiguración de RSTP nos parece suficientemente optimizada para redes campus y difícil y costosa de igualar utilizando protocolos de estado de enlace post-Djikstra. Debe tenerse en cuenta que la estabilidad actual de las redes es resultado de estrategias propietarias desarrolladas para prevenir la inestabilidad de red de los fabricantes [AJY00].

3.4.3 Segmentación en capa dos y jerarquización.

Ante el problema del gran tamaño del dominio conmutado, parece necesaria su segmentación por medios distintos al del direccionamiento IP. Se ha visto también, al exponer los criterios de diseño de redes campus en cuanto a funcionalidad, modularidad de las redes y predecibilidad de las mismas hace, que dichos criterios hacen necesaria una segmentación y una jerarquización, que en las redes campus actuales en la práctica se materializa frecuentemente en tres niveles diferenciados: Núcleo o Troncal, Distribución y Acceso.

Las formas que conocemos para segmentar en capa dos son tres: red superpuesta de capa dos, uso de redes virtuales (VLAN) y uso de regiones de encaminamiento.

La primera forma consiste en crear una red superpuesta [CPV93] con un nivel virtual de bridges de nivel superior mediante encapsulado adicional. La red no se mapea sobre una zona continua necesariamente, pero si es el núcleo, como es el caso de [CPV93] sirve para intercomunicar redes compartimentadas sin crear un dominio único.

La segunda, redes LAN virtuales (VLANs), es la forma estándar de segmentar las redes locales y crear dominios de difusión independientes y de menor tamaño. Hemos visto que la utilización de VLANs requiere configuración detallada y presenta restricciones en su aplicación en toda la red campus por la difusión necesaria de la pertenencia a las VLAN.

En la tercera, definir regiones de encaminamiento separado, se delimitan regiones disjuntas de encaminamiento. Requiere determinar los límites de la red y configurar adecuadamente los bridges que la forman . Es el caso de MSTP, con definición manual de regiones, y de SSTP [Ish+04] para redes Metropolitanas, con particionado automático de regiones, que parece inadecuado para redes campus por su complejidad y menor estabilidad y por no precisarse más que una región en el núcleo.

Se elige la primera diferenciando dos niveles básicos en la red campus, mínimo necesario para establecer una jerarquía: se establece un núcleo y un nivel inferior. El nivel inferior está segmentado en redes más pequeñas interconectadas por el núcleo. Las funcionalidades mínimas e irreductibles de estos niveles consisten en que el núcleo englobará las funciones actuales de los núcleos relacionadas con altas prestaciones, y el nivel inferior las funciones de Acceso. Las funciones intermedias, ligadas normalmente al nivel de Distribución, se asignarán a un nivel u otro durante el proceso de diseño de la solución. Se contemplaría un nivel intermedio de Distribución caso de considerarse necesario.

Con la división en dos niveles queda resuelta la segmentación de la red, compartimentándola a efectos de control, aislamiento de fallos, seguridad, etc.

3.4.3.1 Elección de protocolos de núcleo y nivel inferior

Asimismo se opta por emplear distintos protocolos en el núcleo y en la red inferior, dadas sus grandes diferencias de funcionalidad y objetivos de coste. El núcleo y la red inferior presentan características, requisitos y problemáticas muy diferenciados como se ha descrito en el diseño actual de redes campus, por lo que es presumible que resulten soluciones muy diferentes en ambos niveles.

Protocolos de núcleo. La elección del protocolo para la capa troncal o de núcleo se realiza bajo las siguientes consideraciones:

- Es factible un encaminamiento basado en direcciones MAC entre bridges del núcleo, pero en este caso no puede incluirse encaminamiento a todos los sistemas finales por su gran número. Los bridges del núcleo deben operar de forma separada del resto de bridges para no comprometer la escalabilidad.
- Es factible una solución basada en árboles múltiples inspirada en MSTP, pero deben resolverse sus principales inconvenientes: no optimización de caminos y necesidad de configuración detallada. Presenta la ventaja de menores requisitos de almacenamiento que la solución de encaminamiento. La convergencia puede ser más rápida, a excepción de que se utilice una alta frecuencia de actualización del protocolo de encaminamiento (decenas de milisegundos), con la consiguiente sobrecarga de tráfico (pequeña) y de proceso

En capítulos siguientes de la Tesis se exploran para el núcleo soluciones basadas en árboles múltiples de expansión. Se explorarán alternativas que permitan

autoconfiguración de los árboles múltiples y caminos mínimos y se observa que puede establecerse una instancia de árbol de expansión por bridge de forma automática. Para ello debe ser innegociable el bridge raíz. Ello supone aumentar el número habitual de instancias de árboles de expansión, pero las instancias estandarizadas para MSTP (hasta 64) lo fueron por un criterio relativamente arbitrario, no ligado a la capacidad de proceso requerida en los bridges. Bajo este esquema se diseña el protocolo AMSTP como protocolo derivado de MSTP por simplificación y de ámbito específico para núcleos de redes campus y redes metropolitanas. Posteriormente se tiene noticia de la aplicación de un concepto similar de árboles múltiples en redes L2VPN en el proyecto Global Open Ethernet [Iwa+04].

Protocolo de nivel inferior. En cuanto al nivel inferior, el análisis del estado del arte revela al protocolo STAR [Lui02] como compatible con bridges estándar STP y el más simple de los de su grupo.

Como se ha mencionado más arriba, se desarrolló el protocolo RSTAA como mecanismo de asignación automática de coordenadas. Se observa que STAR presenta algunas carencias en la obtención y estimación de distancias entre bridges, empleando mecanismos indirectos para la deducción de las relaciones topológicas entre nodos, problemas a los que RSTAA es directamente aplicable. Como resultado se desarrolla el protocolo RSJ que combina STAR y RSTAA, con un encaminamiento jerárquico básico basado en vector distancia. Alternativamente, se propone el uso de RSTAR, migración de STAR a RSTP.

Pero, dado el menor coste de los bridges para la red de Acceso respecto a los del Núcleo, dado el predominio del tráfico Cliente-Servidor, y por ser RSTP el protocolo estándar actualmente, se elige como protocolo de nivel inferior para la arquitectura simplificada el protocolo de Árbol de Expansión RSTP, dado que en la mayoría de casos puede no estar justificada la complejidad de utilizar los bridges RSJ además de los bridges ARB en el núcleo con AMSTP. Por el contrario, puede contemplarse para redes campus de tamaño pequeño medio la utilización del protocolo RSJ lo que se propone analizar como trabajo futuro.

3.4.3.2 Mecanismos alternativos a ARP

El mecanismo básico de ARP con difusión debe mantenerse por defecto por ser el único que garantiza la resolución de la dirección en todos los casos, incluido el caso de desplazamiento silencioso de un sistema final. Para reducir el procesado inútil de tráfico ARP por los sistemas finales y mejorar la seguridad, se adopta una solución complementaria basada en servidores ARP en reparto de carga.

Para el problema de la difusión de ARP, la exploración del estado del arte da los conocidos resultados de uso de *proxies* ARP. La exploración de las tecnologías P2P, novedosas en direccionamiento, da resultados negativos por la relativa lentitud de resolución de direcciones y las diferencias en la problemática de la resolución de direcciones en P2P y en redes campus, una operando por encima de IP y normalmente basándose en IP y la segunda operando en la confluencia de IP y MAC, donde se sitúa ARP. No obstante algunas de las técnicas como hash(dirección IP) se eligen para una solución basada en servidores/registradores, dado que servidores deben a la vez garantizar la unicidad y actualidad de la respuesta, se adopta la función de servidor/registrador combinado.

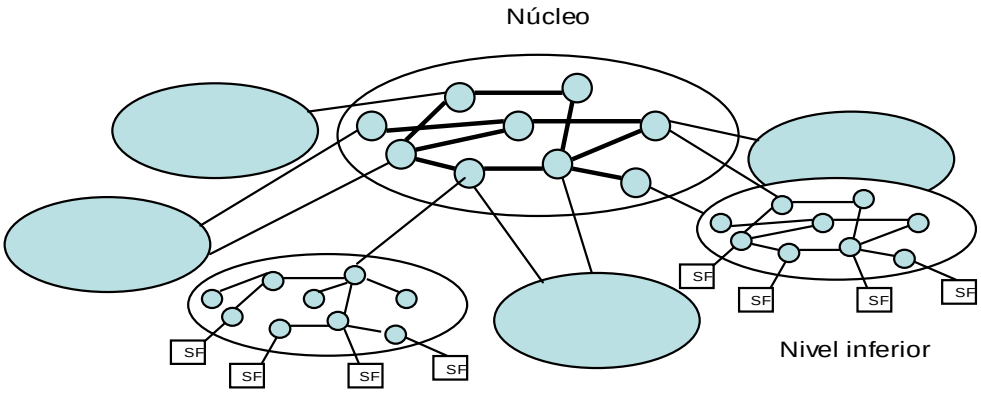


Figura 28. Red jerárquica en dos niveles

3.4.4 Problemas principales de la autoconfiguración.

En las actuales redes campus, el problema principal de configuración se centra en la necesidad de configurar las direcciones IP y diseñar las subredes IP dentro de la red campus. Cuando un sistema final se mueve de un punto a otro de la red campus, su dirección IP deberá modificarse con arreglo a la nueva subred a la que pertenece. Para evitar este problema, se parte de suponer una subred IP única en la red, que independice la dirección IP del punto de conexión en la red.

No obstante, en las redes campus las VLANs realizan una función de compartimentación. Estas VLAN requieren configuración manual, asignando el tráfico según diversos criterios de pertenencia, siendo el más frecuente la pertenencia por puerto de conexión del conmutador, pero no el único, existiendo también la pertenencia a VLANs por dirección MAC (direcciones MAC que deberán estar listadas en un servidor) e incluso por tipo de protocolo de capas superiores. En el ámbito de discusión del IETF [TRILL] este aspecto se separa del problema de autoconfiguración de los Rbridges y no se intenta minimizar.

A juicio del autor, no debe excluirse del ámbito del problema la configuración de VLANs porque existe el peligro de resolver solamente la mitad del problema de la configuración de conmutadores. La autoconfiguración completa de VLANs es probablemente imposible de resolver porque es parte de la definición de las redes virtuales de usuarios, que equivale a una división de la red física en varias redes lógicas. Pero si el objetivo son redes conmutadas autoconfigurables ha de minimizarse el esfuerzo de configuración de las VLAN. Asimismo, pueden resultar incompatibles el uso intensivo de VLANs sin restricciones en las redes campus y el objetivo de autoconfiguración. Una

posición intermedia al respecto parece plausible si desea mantener dicho objetivo, admitiendo algunas restricciones al uso de VLANs, al menos en el núcleo de la red.

3.5 Conclusiones

En este capítulo se ha definido el problema a resolver, y expuesto y valorado los problemas parciales a resolver: difusión de ARP, proliferación de direcciones MAC en las tablas de los bridges, tratamiento de los tráficos multicast en capa 2. Se ha realizado la valoración del Estado del Arte de los protocolos de capa 2 con vistas a su aplicación en el problema planteado, concluyendo en primer lugar que si se encamina mediante direcciones MAC, el dominio debe limitarse de tamaño por razones de escalabilidad. Y en segundo lugar, que si se encamina mediante árboles de expansión, deben optimizarse los caminos y simplificar al máximo la compleja configuración de los árboles de los protocolos como MSTP, manteniendo o mejorando la rápida convergencia de RSTP.

Se ha expuesto el proceso de diseño y las decisiones tomadas para definir la arquitectura, eligiendo un núcleo jerarquizado Ethernet para realizar la compartimentación de la red. Este núcleo, propuesto en [CPV93] con árbol único de expansión (troncal ineficiente), se propone aquí con árboles múltiples autoenraizados, potenciando drásticamente las posibles prestaciones del troncal. Para este núcleo se concibe el nuevo protocolo AMSTP, como protocolo potencialmente óptimo resultado de evolucionar MSTP al problema planteado. Se eligen también los posibles protocolos a emplear para la capa de Acceso con criterios de simplicidad y economía y se propone uno nuevo basado en el empleo de unos identificadores nuevos generados a partir de RSTP. También se ha analizado la problemática que plantea la configuración de las VLANs respecto a los objetivos de autoconfiguración.

Capítulo 4

Arquitectura de red propuesta

En este capitulo se describe la red propuesta que constituye, junto con los protocolos propuestos, el núcleo de la presente Tesis. Tras un resumen de la red propuesta y de los entornos de aplicación, se muestran los escenarios de encaminamiento y los dispositivos utilizados. Los protocolos propuestos se describen en el capítulo siguiente.

4.1 Resumen de la propuesta

El objetivo de diseño planteado es una arquitectura de redes campus de alta capacidad y configuración simple, adecuada para las redes campus futuras de gran tamaño, y que a la vez haga un uso eficiente de la infraestructura de comunicaciones utilizando equipos de coste relativamente reducido. Los tamaños de redes que se contemplan son significativamente superiores a los actuales, dada la previsible proliferación de dispositivos de todo tipo en las redes campus. Consideramos una Red campus de hasta 100.000 dispositivos de los cuales unos 20.000 pueden ser ordenadores convencionales.

Se propone una arquitectura de redes campus en dos niveles y se definen las funcionalidades adicionales de dos tipos de bridges evolucionados (Adaptive Routing Bridges . Se propone un protocolo nuevo para el troncal basado en árboles múltiples de expansión y dos protocolos alternativos para la red inferior. Se proponen también optimizaciones para el trafico multicast y para el problema de la resolución de direcciones en redes campus de gran tamaño.

La red se compone de dos niveles (ver Figura 29) que podríamos denominar como malla de árboles dobles con raíces: un nivel troncal con enlaces ópticos de alta velocidad formado por bridges de troncal (ARB) los cuales encaminan las tramas a través de árboles múltiples, un árbol de expansión distinto para cada bridge frontera. Cada ARB forma parte del troncal y es a la vez Bridge Raíz de un área ó *stub*. El nivel inferior esta formado por la red de cada stub o área, en el cual se encamina por defecto mediante Árbol de Expansión (802.1D) (ver Figura 33). Opcionalmente pueden utilizarse Routing Bridges de un tipo distinto (AB), para aprovechamiento de los enlaces cruzados entre AB de coste inferior a el camino por árbol de expansión. Los ABs ejecutan un protocolo de vector distancia jerárquico basado en coordenadas topológicas jerárquicas. Estas coordenadas son generadas al ejecutar el protocolo de Árbol de Expansión Rápido (RSTP). Mediante el protocolo se obtienen en cada AB rutas alternativas al árbol de expansión que son seleccionadas cuando su coste es menor (ver Figura 39). El tráfico intra-área se encamina entre los bridges 802.1D y los ABs. El tráfico entrante-saliente se encamina vía el ARB correspondiente. Existen ARBs de reserva que asumen la conectividad del stub cuando el ARB activo falla. Estos ARBs pueden funcionar en modo tradicional como bridge raíz de reserva, o en modo modificado como activo-activo, asumiendo uno de ellos toda la red (pasa a ser raíz de toda la red del stub) cuando el otro bridge raíz activo falla. En el modo activo-activo, el stub funciona fragmentado en dos stubs independientes.

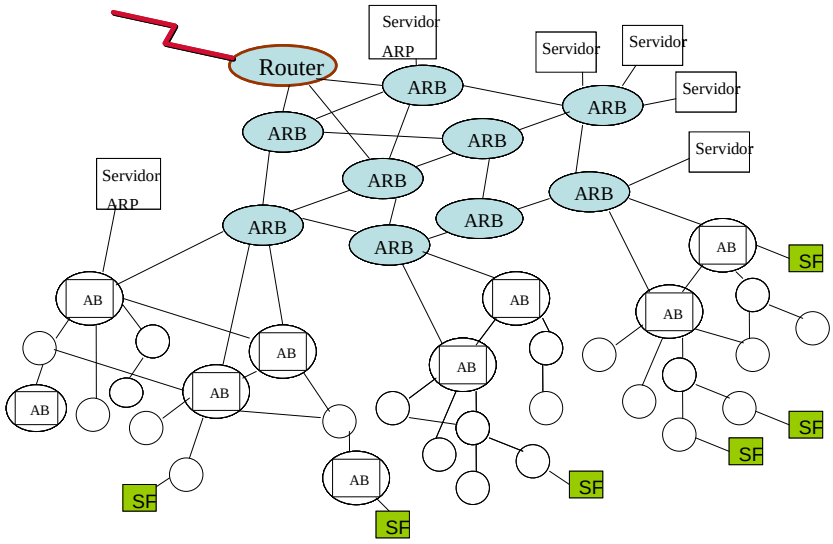


Figura 29. Red campus propuesta (AB opcionales)

Para limitar el tráfico de difusión se utilizan servidores ARP. Para reducir el tráfico multicast de usuario en la infraestructura se recomienda la inclusión de la funcionalidad de inspección IGMP. La topología adoptada combinada con dicha funcionalidad permite una optimización del tráfico multicast en la red troncal.

4.2 Entornos de aplicación

El escenario principal de aplicación de la arquitectura de red propuesta es principalmente el de redes campus de media y gran capacidad. Adicionalmente, la arquitectura propuesta es aplicable en troncales de Redes de Area Metropolitana. A continuación se describen brevemente ambos.

El principal entorno de aplicación es el descrito en los requisitos en el capítulo 3 que es el de redes campus de medio y gran tamaño como las que pueden esperarse en el inmediato futuro en empresas, universidades, etc. Estas redes podrían tener hasta 100.000 sistemas finales de diferente entidad, siendo la mayoría de ellos dispositivos y no PCs en una proporción aproximada de 80/20.

Pero los conceptos que incorpora la propuesta para redes campus descrita en la presente Tesis son aplicables a otros escenarios de forma más conservadora o digamos *evolutiva* respecto a las redes existentes actualmente. Como ejemplo de escenario adicional que consideramos adecuado para la propuesta descrita en la presente Tesis, en particular para el protocolo AMSTP, son los troncales ópticos de área metropolitana. Un caso se describe de forma simplificada en la Figura 8. Este escenario de aplicación es similar al propuesto por Duato (utilizando encaminamiento de direcciones MAC entre los bridges de troncal) [GDS03]. En nuestro caso se emplea por el contrario el protocolo AMSTP para reenviar los paquetes mediante árboles de expansión múltiples desde cada Bridge Frontera Origen en el troncal hasta el Bridge Frontera Destino [IGA04]. En este entorno de aplicación, el troncal está formado por un determinado número de conmutadores interconectados por enlaces de 10 Gigabit Ethernet. Cada conmutador recibe el tráfico de las LAN mediante enlaces de 1 Gbps. Típicamente las LANs estarán conectadas al troncal mediante routers. En estas condiciones, el número de direcciones MAC visible por los bridges es bastante reducido (típicamente del orden de centenares) y el así conocido problema de proliferación explosiva de direcciones MAC (*MAC address explosion problem*) no se produce ya que los routers separan los dominios de conmutación y los bridges del troncal solamente aprenderán las direcciones MAC de otros bridges y de los routers terminales.

En este entorno de aplicación no es necesario el uso de servidores ARP para la contención de la difusión de paquetes ARP.

4.3 Escenario de encaminamiento

A continuación describimos un ejemplo del proceso de encaminamiento en la red campus. Los supuestos que hacemos en la red son los siguientes:

-El troncal está formado por bridges ARBs. Los bridges no ARBs conectados a los ARBs no participan en el protocolo AMSTP, operando de forma estándar IEEE 802.1D en el reenvío de tramas.

-En las áreas del nivel inferior, bajo cada ARB, coexisten bridges AB y bridges estándar en las regiones bajo cada ARB. El nivel inferior está formado por áreas

independientes cada una bajo un ARB, en las que se ejecuta un protocolo de Árbol de Expansión complementado con un protocolo que mejora la utilización de la red empleando enlaces cruzados. Los dispositivos que forman el nivel inferior son bridges estándar 802.1D y bridges AB. Los bridges AB son los encargados de ejecutar el protocolo complementario (RSTAR o RSJ). El bridge raíz del área es siempre un ARB para realizar la función de acceso al troncal, por lo que deben preverse bridges que puedan asumir la función ARB en caso de fallo del ARB principal.

-El encaminamiento intra-área ARB utiliza el protocolo RSJ o el RSTAR y el encaminamiento inter-área el protocolo AMSTP. En el primer caso hay doble encapsulado RSJ, realizado en los ABs y en el segundo se reencapsula en el ARB de área.

4.3.1 Escenario de encaminamiento completo

Veamos un ejemplo del proceso genérico de encaminamiento en la red campus. En la red de la , consideremos el caso de un sistema final conectado a un bridge 802.1D que debe enviar paquetes a otro sistema final. Describiremos primero la resolución ARP del destino y después el encaminamiento por la red campus. La resolución ARP se realiza como sigue (ver Figura 30):

- El sistema final envía un paquete ARP en difusión (*broadcast)*. El AB agente (primer bridge RSJ en el árbol de expansión hacia arriba) lo intercepta,
- Calcula hash(IP) de las direcciones IP origen y destino de la trama ARP.
- De *hash(IPorigen)* obtiene *servidor_ ARP_registrador.*
- De *hash(IPdestino)* obtiene *servidor ARP_ resolucion_ARP*
- Segun el resultado de dichos hash obtiene la direcciones unicast de los servidores ARP encargados de dichas direcciones IP a efectos de consulta y registro.
- Encapsula la trama recibida con cabecera adicional de capa 2 y lo envia a los dos servidores ARP:
    - Con dirección MAC destino unicast del servidor ARP que le correponde y su dirección MAC (AB origen) (resolución ARP, servidor ARP en rol de consultado)
    - Con dirección destino (unicast) MAC del servidor ARP que le corresponde a la IP del que pregunta (para registrar al sistema final que pregunta)
        - El servidor ARP (en rol de *registrador* de la IP origen) recibe la trama doblemente encapsulada y registra o renueva la tupla:
            - *MAC sistema final, MAC bridge agente AB, coordenada RSJ AB, ARB del árbol.*
        - El servidor ARP *consultado* contesta al paquete ARP enviando al AB una trama con doble encapsulado:
        - *dirección MAC AB , dirección MAC sistema final , IP resuelta, ARB del destino, AB del destino, MAC sistema final*
        - El AB recibe la trama de *ARP_response*, la desencapsula y la reenvía al sistema final en el formato estándar.

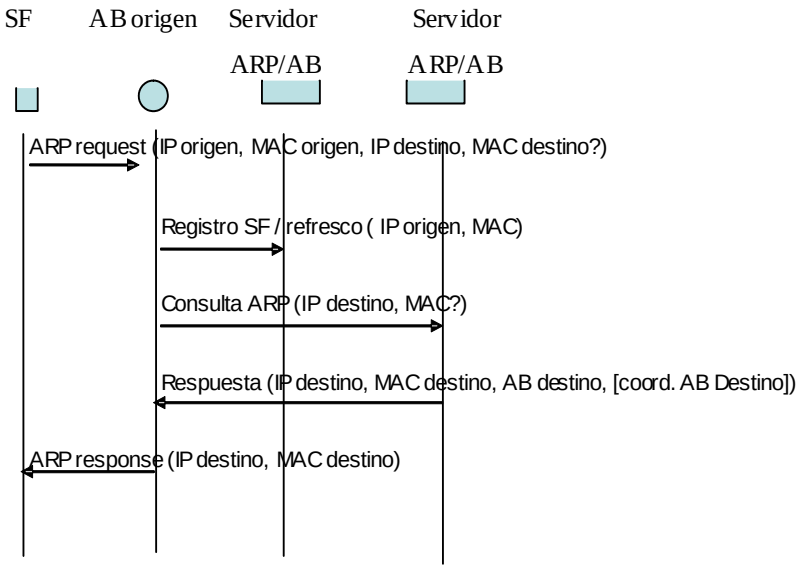


Figura 30. Proceso de ARP/AB con servidores

- El sistema final envía su primer paquete IP hacia el sistema final destino.
  - El AB lo intercepta,
  - Si el AB conoce la MAC del SF destino, lo encapsula con direcciones *MAC AB destino* y *MAC AB origen*, incluyendo la coordenada RSJ del AB destino.
  - Si el AB no conoce la MAC del SF destino, inunda como un bridge estándar, sin encapsular.
- Una vez que la trama encapsulada con dirección destino la del AB destino sale del AB existen dos casos a considerar:
- El AB destino está en la misma área bajo del ARB (encaminamiento intra-área)
- El AB destino está en un área distinta, bajo otro ARB (encaminamiento inter-área)

El registro por parte del AB contempla otros casos adicionales a la petición ARP, como el ARP *response* para el caso de servidores que no realizan ARP, o en general tramas Ethernet con dirección MAC origen que no aparezcan en la caché del AB agente.

4.3.1.1 Encaminamiento intra-área

La trama encapsulada con el AB destino y el AB origen se encamina según el protocolo intra-área elegido (RSTP por defecto, RSJ o RSTAR opcionalmente).
Si se utiliza RSTP o, utilizando RSTAR no existe camino alternativo mejor, la trama se envía por el Árbol de expansión sin encapsular.

Esto quiere decir que el AB origen busca en las tablas si tiene una ruta para ese AB destino mejor que el de via Árbol de Expansión, en cuyo caso encapsula la trama con dirección MAC destino la dirección del AB siguiente salto, la de la tabla de rutas.

La trama encapsulada llega al AB de siguiente salto donde se consulta la tabla de rutas y se reemplaza por el AB siguiente salto hasta llegar al AB destino, que es el agente

del sistema final destino. En el caso de utilizar encaminamiento con protocolo RSTAR, los sistemas finales difunden listas de sistemas finales entre ellos y en cada salto se decide el bridge AB de siguiente salto por sistema final destino. En caso de encaminar con el protocolo RSJ, la coordenada RSJ del AB destino se incluye en la trama, y es lo que se usa para buscar en las tablas de encaminamiento si hay camino alternativo (cross link) mejor que el árbol de expansión. En cada salto igualmente se cambia la MAC por la MAC del AB de siguiente salto, para evitar la posible duplicación de la trama por los bridges 802.1D de la capa de acceso [Lui02].

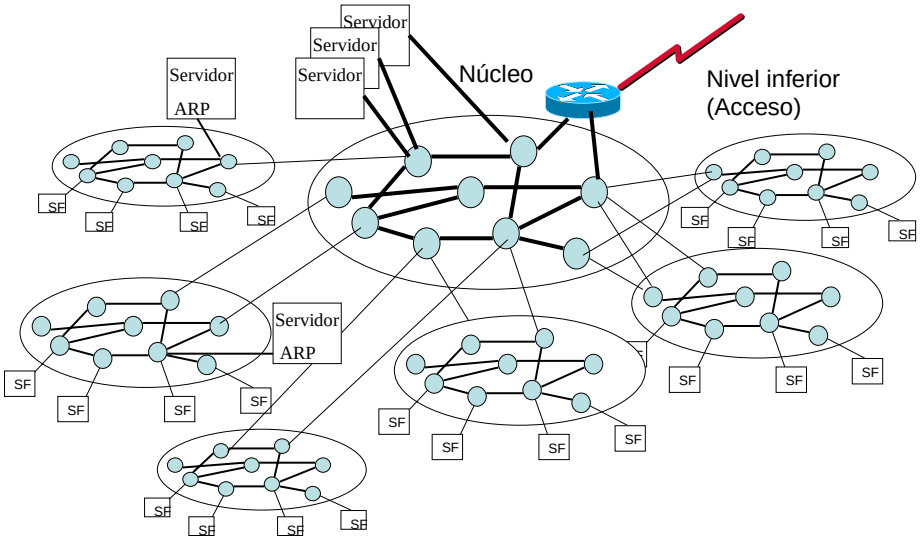


Figura 31. Esquema general red en dos niveles.

4.3.1.2 Encaminamiento ínter-área

En la Figura 32 se muestra el escenario de encaminamiento ínter área. La trama encapsulada con el AB destino/AB origen es encaminada dentro del área como cualquier otra. Como la dirección del AB destino no es conocida por el protocolo de encaminamiento de área, la trama se encamina por el árbol de expansión llegando hasta el bridge raíz (ARB). El ARB consulta su tabla de AB-ARBs, obteniendo el ARB cabecera del área del AB destino, reencapsula la trama con *ARB destino-ARB origen* (eliminando la cabecera recibida *AB_destino-AB_origen*) y se envía por el árbol de expansión correspondiente. Como en este árbol de expansión los bridges han aprendido las direcciones de los ARBs, solamente se enviará por los enlaces de la rama que une el ARB origen con el ARB destino. Si no lo conociera, se realiza difusión a todos los ARB.

El ARB destino recibe la trama y la desencapsula, enviándola por el árbol de expansión por el puerto por donde haya sido aprendida esa dirección de sistema final (o inundando por el árbol si hubiera caducado en su caché). La trama se transmite desde el ARB destino, que es bridge raíz del árbol, directamente hasta el sistema final destino tras atravesar el AB agente del SF.

Recapitulando, el servidor ARP devuelve el AB destino al AB origen con la contestación al ARP, el AB encapsula con AB destino-AB origen y envía al ARB por el árbol, mediante el AB destino el ARB origen resuelve el ARB destino (los ARBs se intercambian periódicamente sus ABs (son pocos) y cambia (desencapsula y reencapsula) el encapsulado de *AB-AB a ARB-ARB.* El ARB destino desencapsula y envía directamente la trama desencapsulada directamente por el árbol de expansión (filtrada a su rama solamente por aprendizaje).

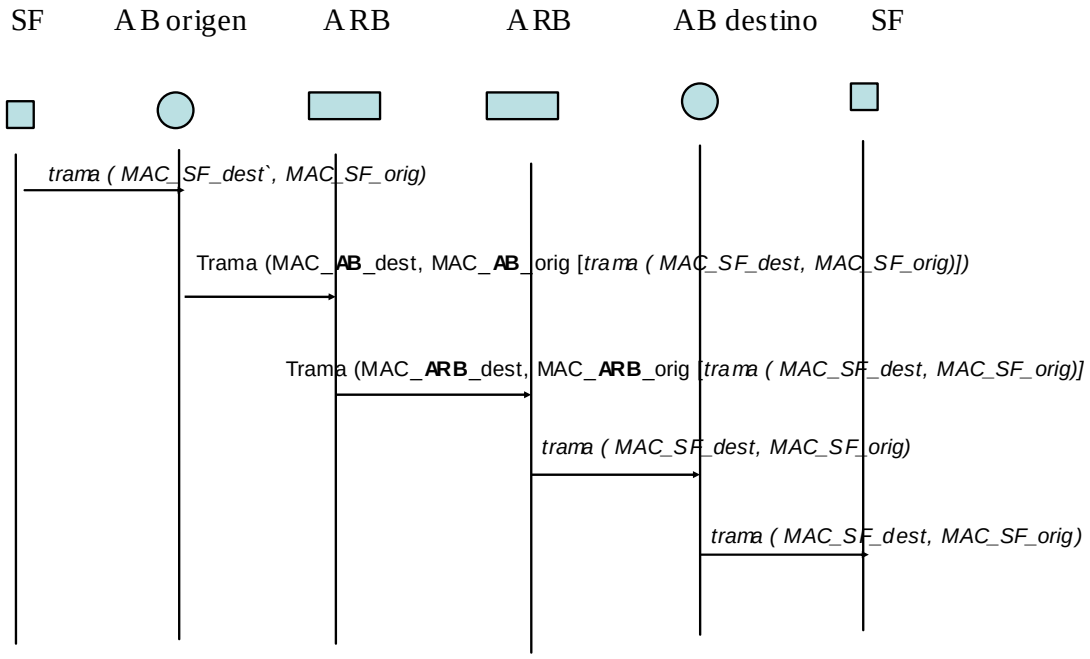


Figura 32. Encaminamiento de trama en escenario completo

4.3.2 Escenario de red simplificado

Cuando el tamaño de la red campus no es muy grande y se desea la máxima compatibilidad de equipos en el nivel inferior, la propuesta se puede simplificar mucho. (Ver Figura 33). La simplificación se realiza en la red inferior principalmente. Los bridges de tipo AB no se utilizan, solamente existen bridges 802.1D estándar y ARBs, realizando el bridge ARB raíz del correspondiente árbol de la red inferior todas las funciones no estándar residuales de los bridges AB. La funcionalidad del protocolo AMSTP se mantiene y los ARB asumen las funciones de registro y resolución de ARP ante los servidores ARP.

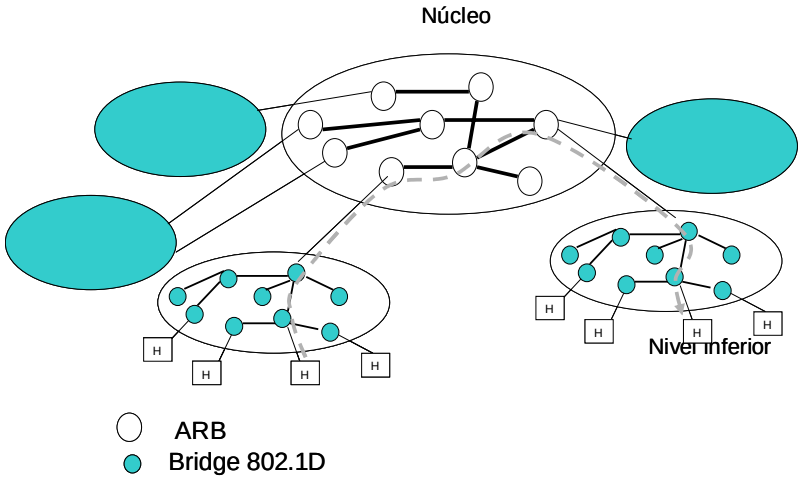


Figura 33. Escenario de red simplificado.

4.3.2.1 ARP

En la Figura 34 se muestra el proceso de resolución y registro en servidores ARP en el escenario simplificado. El proceso es el siguiente:

- El sistema final envía la trama ARP en difusión:
- La trama ARP se difunde por toda la rama del árbol de expansión llegando hasta el bridge raíz (ARB). (Si el sistema final destino del ARP está conectado en esa misma rama, responderá a la trama ARP de forma estándar).
- El ARB reconoce que se trata de una trama ARP. La trama ARP no es redifundida por el ARB hacia otras ramas del árbol de expansión.
- El ARB calcula *hash(IP_SF_destino)* y *hash(IP_origen)*, obteniendo las direcciones de los servidores ARP/ARB correspondientes y envía sendos paquetes para el registro del SF origen y el ARP request del SF destino a los servidores ARP/ARB.
- El servidor ARP/ARB contesta incluyendo en el paquete de respuesta las direcciones MAC del SF destino y del ARB agente del SF destino. El ARB guarda en caché ésta información.

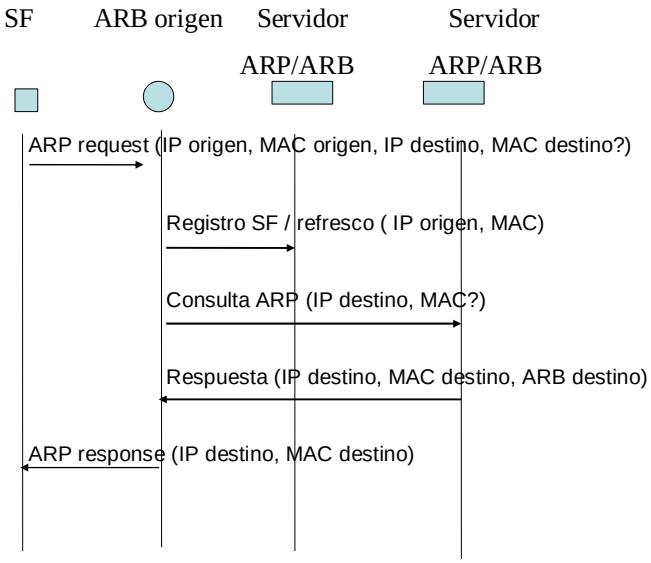


Figura 34. ARP con servidores en escenario simplificado

4.3.2.2 Encaminamiento de trama

En la Figura 35 se muestra la secuencia de encaminamiento de una trama entre sistemas finales (SF).

- El ARB actúa como Bridge Agente y encapsula la trama recibida con la cabecera adicional de capa 2 que contiene las direcciones MAC del puerto designado del ARB origen y del puerto del bridge ARB destino. Esta dirección se conoce porque es difundida por el protocolo AMSTP desde cada ARB raíz en sentido descendente y va incluida en la BPDU AMSTP.
- El ARB origen envía la trama por el puerto raíz correspondiente a la instancia de árbol de expansión cuyo raíz es el bridge destino.

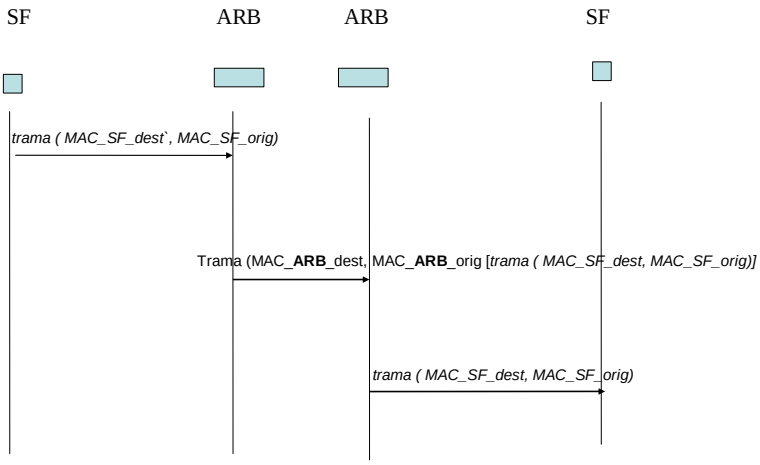


Figura 35. Envío de trama en escenario simplificado.

4.4 Dispositivos

Se describen a continuación los tipos de dispositivos utilizados en la red propuesta. Los dispositivos adicionales propuestos son: ARBs, ABs y servidores ARP. Se describe el encapsulado adicional de capa 2 que realizan los AB y ARB. El empleo de los ABs es opcional, no se utilizan cuando el protocolo del nivel inferior es RSTP. Los servidores ARP son también opcionales existiendo diferentes alternativas a su función. Las tablas y estructuras de datos en los dispositivos se describen en el capítulo siguiente, tras la descripción de los protocolos para una mejor comprensión.

4.4.1 ARBs

En la red propuesta, por tratarse de una subred IP única, la posición y función de los encaminadores en la red campus es solamente la de pasarela hacia las redes externas y no la habitual de dividir la red campus en subredes. Esta función de compartimentación interna de la red campus es realizada por los ARBs. Los ARB constituyen el núcleo de la red o troncal como nivel jerárquico superior y asimismo realizan la función de bridge frontera de entrada y salida (*ingress/egress backbone bridge)* de las tramas desde/hacia el nivel inferior de la red. Los ARB actúan como bisagra de la red, dado que actúan como bridge raíz tanto en el nivel inferior como en el troncal en sus respectivas instancias de árbol. Como se describe más adelante, los ARB tienen una funcionalidad en los puertos del troncal y otra en los conectados a la red inferior. Los ARB ejecutan el protocolo AMSTP para las funciones del troncal y el protocolo RSTP/STP a efectos de árbol de expansión en el nivel inferior.

4.4.1.1 Los ARBs como bridges dobles o virtuales.

Los ARBs, para realizar su función de enlace entre los dos niveles de la red Troncal y Distribución-Acceso y al mismo tiempo las funciones de encaminamiento en el nivel superior (troncal), se comportan como bridges duales, es decir el bridge se puede considerar formado por dos bridges unidos internamente e intercomunicados, como se muestra en la Figura 36. Los puertos del bridge conectados al troncal, tambien llamados *puertos de troncal*, forman parte de un bridge que ejecuta el protocolo AMSTP. A los puertos del bridge unidos al nivel inferior de la red los denominamos *puertos de distribución*, y funcionan entre sí como los puertos de un bridge estándar. Lo mismo sucede entre los puertos de troncal. En la comunicación entre ambos bridges se produce el encapsulado y desencapsulado de forma análoga al de *bridge virtual* propuesto en [Cas+94], que es la propuesta más temprana que conoce el autor de utilización de encapsulado adicional entre bridges para construir un nivel superior de bridges o bridge virtual, posteriormente generalizado en [Tou+03].

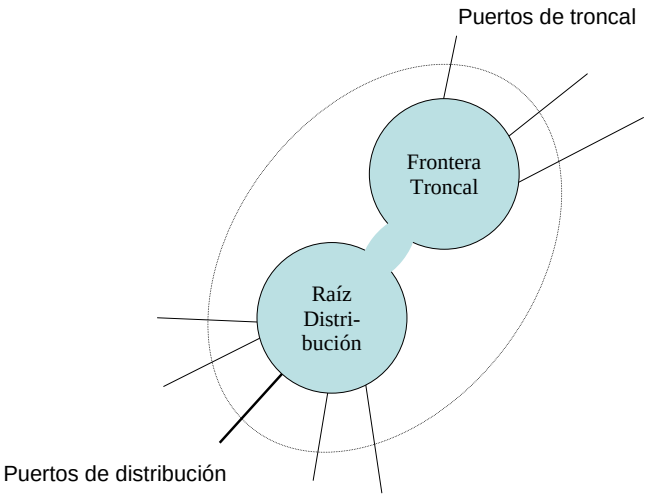


Figura 36. Componentes de un ARB

Cada ARB se comporta en el nivel troncal como bridge raíz no negociable de su propio árbol de distribución AMSTP y en el nivel inferior como bridge raíz RSTP.

Los puertos del lado 802.1D (estándar) del ARB, por configuración o porque detectan BPDUs 802.1D estándar, distribuyen las BPDUs estándar RSTP y operan anunciando al ARB como bridge raíz, los puertos del ARB conectados a otros ARBs, mediante enlaces punto a punto, ejecutan AMSTP (que a su vez requiere la ejecución de otro RSTP) y distribuyen AMSTP BPDUs (o bien BPDUs RSTP si escuchan tramas RSTP por razones de compatibilidad). Este funcionamiento dual del ARB es similar al de los bridges que emplean doble encapsulado en Ethernet (VLAN sobre VLAN (*Q-in-Q)*, o MAC sobre MAC. Las BPDUs emitidas por el ARB son distintas según el enlace pertenezca a una parte del bridge o a la otra. En los enlaces hacia la red inferior se anuncia como raíz, en los enlaces hacia otros ARBs será elegido el ARB de menor Bridge ID.

La conectividad a efectos de BPDUs para construir los árboles de expansión STP (bridge raíz, costes de caminos) no se difunde de una parte del ARB a otra. Los estados de puerto tampoco, por lo que entre los enlaces hacia el troncal se difunden las BPDUs con información de puerto raíz, puertos designados, etc., del árbol superior y lo mismo independientemente en los enlaces hacia abajo para el nivel inferior.

Los puertos pertenecen a un árbol u otro según su configuración por defecto, su configuración explícita y según las BPDUs que escuchan. Mediante un mecanismo como el de migración de protocolos de puertos de 802.1D [802.1D04], difunden BPDUs compatibles con las recibidas. En los puertos en los que los enlaces no sean punto a punto el protocolo ejecutado es STP, dado que tanto RSTP como AMSTP requieren enlaces punto a punto. Esta característica se lee en una variable de la MIB del protocolo.

4.4.1.2 Bridge IDs de un ARB.

En este epígrafe se hacen unas consideraciones de tipo más bien práctico sobre la implementación de las identidades de bridge en el ARB. Se ha descrito más arriba la

doble funcionalidad de un ARB. La cuestión práctica que surge es si se han de utilizar dos identidades (Bridge ID) distintas para ambos componentes del ARB o si una sola puede bastar. Un bridge ARB participa en dos árboles de expansión básicos y en múltiples instancias de árboles. Respecto a los árboles de expansión básicos, el ARB participa como un elemento más en el árbol de expansión del troncal (establecido mediante el protocolo RSTP) y como raíz en el árbol de expansión de la red inferior.

Respecto a las instancias múltiples de árboles, el ARB participa como raíz de una instancia (su árbol de expansión, ése ARB como destino) y como nodo normal en las N-1 instancias de árbol restantes correspondientes a los N-1 nodos.

Por simplicidad de gestión, el ARB tiene una única identidad de Bridge (Bridge ID) para ambos Árboles de Expansión (IST en red superior y ST en inferior). Puede ser conveniente en ciertos aspectos el que dichas Bridge ID sean completamente distintas, configuradas de forma totalmente independiente, o bien difieran solamente en parte, concretamente en la prioridad configurable (por ejemplo un rango de prioridades para el núcleo y otro rango para los niveles inferiores). La Bridge ID difundida por el ARB en los puertos de distribución puede ser la misma o distinta que la difundida por los puertos de troncal. Parece conveniente que haya dos Bridges IDs, pero puede ser práctico emplear una única dirección MAC y prioridades distintas (es decir, lo que es distinto es la parte de prioridad de la Bridge ID). En este caso podría formarse mediante una parte configurable (funcionamiento ARB habilitado más prioridad en red superior, 16 bits) y una fija, consistente en la dirección MAC del bridge normal 48 bits). Puede ser simplemente otra prioridad de 16 bits y que sea alterado su valor por habilitar o no el funcionamiento como ARB. En todo caso debe tenerse en cuenta que la bridge ID utilizada a nivel de troncal es la usada para elegir bridge raíz del troncal por el protocolo RSTP del troncal, aspecto que debe estar controlado mediante la configuración de la prioridad del ARB previsto como raíz en el troncal.

En cuanto a la bridge ID para el troncal, los protocolos con instancias múltiples de árbol es necesario a veces utilizar identidades de bridge adicionales, una en cada bridge por cada instancia. A este respecto y como es bien conocido, en el estándar 802.1D la bridge ID se compone de dos octetos de prioridad que se configuran y seis octetos de la dirección MAC utilizada como Bridge ID. En el caso de utilizar VLANs, cada instancia de VLAN requiere que sea configurada una Bridge ID distinta, lo cual causa una necesidad excesiva de direcciones MAC, que se puede evitar mediante el uso combinado de la dirección MAC y la identidad de VLAN para generar una bridge ID distinta que es utilizada en la negociación de cada instancia. En cuanto a las instancias de árboles múltiples AMSTP, un ARB es raíz de una instancia y participa en las demás, la identidad de bridge no se utiliza para elegir bridge raíz de las instancias múltiples porque el bridge raíz no es negociable en AMSTP, todos los ARB frontera (con puertos de distribución activos) crean una instancia de árbol AMSTP. Por ello en AMSTP la prioridad de cada bridge no influye en la elección de bridges de los árboles múltiples de expansión, pero sigue siendo necesario crear o utilizar Bridge ID s distintas del mismo bridge para representarlo en cada instancia de árbol. Un posible mecanismo es utilizar, de forma similar a el caso de VLANs/MSTP, 12 de los 16 bits de prioridad asignándoles el resultado de *hash (MAC ARB)* raíz de la correspondiente instancia de árbol. De esta forma, cada ARB usará una Bridge ID en cada instancia de árbol, de la forma:

Bridge_ID (este_ARB, instancia_ARBj) = unir (hash$_{16bit}$ (MAC ARBj), MAC_este_ARB)

En el árbol del nivel superior (troncal), el bridge raíz se elige de acuerdo con el protocolo RSTP estándar, por lo que el que tenga menor (mejor) prioridad configurada será el elegido como bridge raíz. La prioridad de los bridges se configura de forma sencilla mediante comando en los bridges estándar, resultando en la modificación de los dos octetos más significativos de la bridge ID que preceden a los 48 bit a dirección MAC del bridge y forman conjuntamente los 64 bits de la Bridge ID.

4.4.1.3 Comunicación entre ARBs y entre ARBs y ABs

Los protocolos ejecutados requieren comunicación entre los ARBs y de los ARBs con los ABs. Cada protocolo se describe por separado en una sección de este capítulo. Describimos a continuación las direcciones empleadas para la comunicación de los ABs y ARBs:

-Se utiliza una dirección de grupo multicast para los ARBs y otra para los ABs. Además cada ARB y AB emplea las direcciones MAC unicast correspondientes a cada uno de sus puertos. Los ARBs se comunican con los ABs principalmente por su función de bridge raíz del árbol de Expansión y su función de pasarela con la red troncal.

-Normalmente un ARB dialoga con los demás ARBs por sus puertos de troncal y con los ABs por sus puertos de distribución. Un AB dialoga con los ABs por todos sus puertos designados y raíz y con el ARB de su área por su puerto raíz solamente.

Cada ARB dialoga (escucha, procesa y contesta) en las direcciones multicast ARB:
-En la dirección multicast ARB: el intercambio de listas de ABs entre ARBs.
-En la dirección multicast de BPDUs estándar (STP, RSTP) para las BPDUs de dichos protocolos y las BPDUs AMSTP.

4.4.1.4 Aspectos de formación de los árboles de expansión en ARBs

Los bridges ARBs se configuran por defecto con un valor de prioridad inferior al valor de prioridad por defecto de los bridges estándar, a fin de evitar que algún bridge 802.1D de su red pueda ser elegido bridge raíz del troncal de ARBs. La elección de un ARB u otro como bridge raíz del troncal ARB depende solamente de la prioridad y bridge ID de los ARBs que estén conectados en dicha red

4.4.1.5 Modos de funcionamiento de los ARBs.

Los ARB pueden funcionar en modo activo-activo conectando ambos ARB al troncal y cada uno siendo bridge raíz de árboles de expansión disjuntos entre sí. Si falla uno de ellos, los dos árboles se unen en uno, quedando como bridge raíz el bridge ARB del par que no ha fallado. El modo activo-activo requiere pequeños cambios en el algoritmo de elección de bridges raíz para que ambos bridges utilicen una misma bridge ID a efectos de ser elegidos simultáneamente como bridge raíz. Los enlaces que conecten los bridges de ambos árboles de expansión resultantes dependientes de los dos ARBs respectivos, serán bloqueados por el protocolo, creándose dos árboles independientes.

Los ARBs pueden alternativamente utilizarse en modo activo-reserva. En este modo de funcionamiento tanto el ARB activo como el ARB reserva tienen conectividad al troncal, pero solamente el ARB activo es elegido como bridge raíz por todos los bridges del árbol por tener configurada su prioridad adecuadamente. El ARB en modo reserva funciona como bridge AB del árbol al que pertenece, cursando tráfico intra-área, hasta el

momento en que el bridge ARB activo falle y el bridge reserva asuma el papel de bridge ARB, raíz de la red inferior. Los tiempos de reconfiguración en el modo activo-reserva son superiores a los del modo activo-activo por la reinicialización del AB que pasa a operar como ARB. La implementación de la Figura 37 para la configuración activo-activo emplea bridges AB adicionales. En este ejemplo se utiliza el ARB con un solo puerto de distribución de forma que se facilita la conmutación de toda la red bajo el AB de un ARB al ARB superviviente. En caso de fallo de un ARB, el enlace de reserva correspondiente que une el AB al otro ARB se reactiva quedando dicho ARB a cargo de toda la red inferior formada por las áreas A y B, las cuales quedan de hecho unidas bajo el mismo ARB raíz. Los AB deben implementar la modificación necesaria del protocolo RSTP para que permitan elegir distinto ARB como bridge raíz cuando ambos ARB están disponibles. La modificación consiste en comparar solamente la prioridad al elegir bridge raíz, ignorando la identidad de bridge.

4.4.1.6 Configuración Activo-Activo mediante semi-ARBs

Para obtener redundancia en la conexión a la red troncal, una posibilidad de optimización consiste en implementar la pareja de ARBs mostrada en la Figura 37 como dos semi-ARBs. El ARB presentará un modo de fallo por mitades. Cada semi-ARB consta de la mitad de los puertos del ARB total. Para hacer posible el funcionamiento independiente de un semi-ARB ante fallo del otro, las máquinas de estados RSTP de dichos puertos y la máquina RSTP de selección de roles de puerto (ver más adelante máquinas de estados RSTP en Figura 4) se deben implementar como dos mitades interconectadas, cada una de las cuales controla la mitad de los puertos. El desarrollo de este concepto u otros alternativos está fuera del objetivo de esta Tesis.

Todos los bridges situados en el árbol de expansión debajo de este ARB formado por dos semi-ARBs pueden ser estándar 802.1D porque dependen de un mismo bridge raíz y no precisan modificación alguna en el protocolo de árbol de expansión.

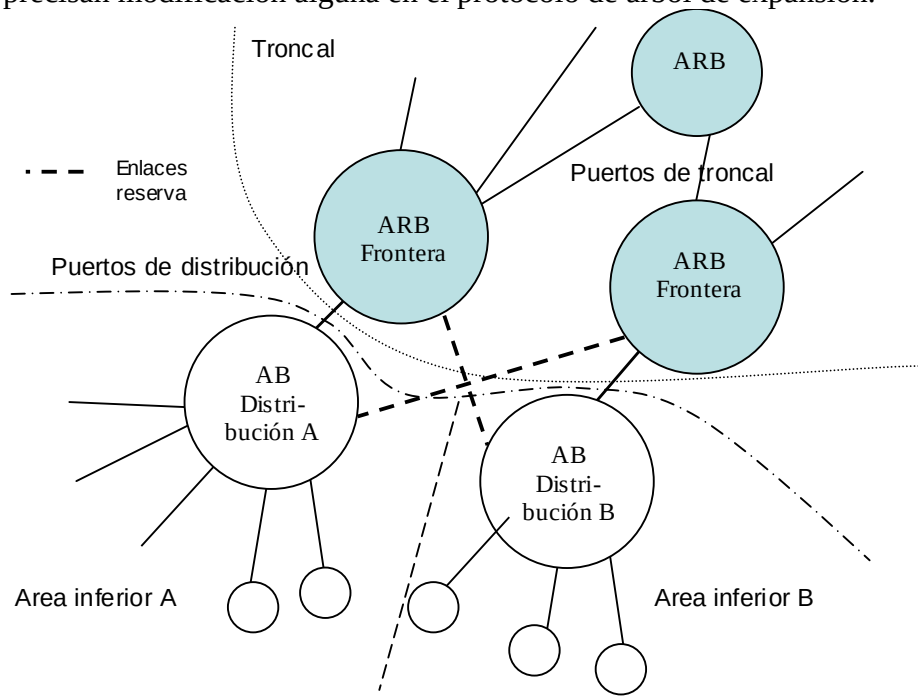


Figura 37. Funcionamiento ARB activo-activo.

4.4.2 Adaptive Bridges (AB)

Los AB son los bridges de funcionalidad extendida del nivel inferior de la red. Su utilización es opcional, no se requieren en el escenario simplificado descrito. Su función se ha explicado al describir el encaminamiento, son los bridges agentes de los Sistemas Finales, resuelven el ARP ayudados por los servidores ARP, encapsulando y encaminando las tramas doblemente encapsuladas en la red inferior hasta otro AB del área ó hasta el ARB (bridge raíz). Los ABs deben detectar determinados protocolos en la trama Ethernet (como ARP). Los AB, además de las funciones estándar de un bridge 802.1D, ejecutan el protocolo de encaminamiento de área elegido (RSTAR o RSJ) que permite utilizar caminos alternativos al árbol de expansión (enlaces cruzados), de menor coste. Si hay varios interfaces de red en un sistema final, éstos se comportan en la red campus como sistemas finales distintos, al no existir un identificador único de host, (tipo HIP por ejemplo). Cada interfaz estará asociada a un AB designado, solamente a uno, el primero según asciende por el árbol de expansión.

4.4.3 Bridges 802.1D

Denominamos bridges 802.1D a los bridges estándar, normalmente situados en el área inferior, que ejecutan RSTP o STP. Si todos los enlaces de la red entre bridges son punto a punto, cuestión recomendable por muchas razones incluida la seguridad, puede ejecutarse el protocolo RSTP. Si existen enlaces compartidos el protocolo de árbol de expansión es STP, de convergencia más lenta, y el protocolo de encaminamiento en el área sería el protocolo original STAR [Lui02]. Los bridges estándar conectados al troncal operan de forma estándar con los ARBs, pero no participan del protocolo AMSTP.

Los bridges estándar se conectan al árbol de expansión existente en el área de forma estándar, por lo que su prioridad no debe configurarse de forma que puedan ser elegidos como bridge raíz. Cabe imaginar que algunos de estos bridges estén configurados para ejecutar RSTP y adicionalmente MSTP. En este último caso pertenecerán a regiones MST distintas al área que nos ocupa, por lo que las BPDUs MSTP (por tener limitación de saltos) no interfieren con el área de los ARBs.

4.4.4 Encaminadores

Los encaminadores de la red campus se conectan directamente a los ARBs y su función en la red campus, si la compartimentación es realizada por los ARBs, es la de actuar como pasarelas hacia la red exterior a la red campus. Es razonable, factible y conveniente en ocasiones que los encaminadores incorporen funcionalidad de ARB opcionalmente para obtener equipos más compactos.

4.4.5 Encapsulado en ARBs y ABs

El mecanismo básico de los Routing Bridges y en general de todo tipo de redes virtuales o superpuestas (*overlays)* es el encapsulado adicional. Tanto los ARBs como los ABs utilizan un mecanismo de encapsulado adicional de capa 2 para comunicarse entre

ellos coexistiendo con bridges 802.1D, realizando un procesado diferente a las tramas encapsulada. El esquema de encapsulado puede verse en la Figura 38. Aunque tanto los ARB como los AB utilizan encapsulado de capa 2 existen diferencias en su procesamiento derivadas principalmente de que mientras que AMSTP encamina por árboles de expansión, RSTAR y RSJ encaminan por Vectores Distancia salto a salto.

4.4.5.1 Nivel inferior

El nivel inferior corresponde al encapsulado/desencapsulado realizado por los AB. Los ARB realizan también este encapsulado/desencapsulado con el tráfico de entrada y salida del troncal hacia el nivel inferior.

BPDUs en nivel inferior

En la red inferior se utiliza el encapsulado propuesto por el protocolo STAR [Lui02]. Tanto los protocolos RSTAR como RSJ del nivel inferior, utilizan RBPDUs de formato similar a las BPDUS STP, con su propio identificador de protocolo, distinto al de STP . Los bridges RSTAR o RSJ saben procesar tanto las BPDUs estándar STP/RSTP como las SBPDUs (RSTAR/RSJ). Como un bridge estándar podría descartar estas RBPDU por desconocer el protocolo, para evitar que las RBPDUs sean descartadas se utiliza encapsulado de manera similar a STAR y se transmiten como tramas de datos encapsuladas con la dirección MAC del AB origen y con la dirección MAC destino del AB (es decir la dirección del AB siguiente salto, dado que en los ABs se encamina salto a salto). También se utilizan como dirección destino la dirección multicast *"todos los bridges AB"*.

Tramas de datos

Las tramas de datos pueden ser encapsuladas o no. Cuando se encapsulan, se añaden como cabecera la dirección MAC del bridge RSTAR/RSJ destino y origen. Estas tramas tienen formato Ethernet con un Ethertype específico que permite a los bridges RSTAR/RSJ distinguir una trama encapsulada de una trama normal.

Las tramas de datos son reenviadas normalmente por los bridges estándar desde el bridge RSTAR/RSJ hasta el bridge RSTAR/RSJ destino. En cada salto la cabecera exterior de la trama se modifica con las nuevas direcciones MAC origen y destino del salto correspondiente.

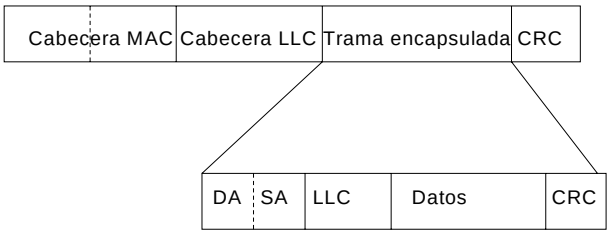


Figura 38. Formato genérico de trama encapsulada

4.4.5.2 Nivel troncal

En el nivel troncal el protocolo utilizado es AMSTP, que coexiste con RSTP y STP. En el nivel troncal las direcciones origen y destino MAC de la cabecera exterior de las tramas no se cambian al atravesar cada bridge ARB, dado que se encamina por instancias de árbol.

BPDUs

En la descripción del protocolo AMSTP indicamos los mecanismos de compatibilidad de BPDUs y distinción de BPDUs propuestos. En este caso se utiliza la misma identidad de protocolo y diferente variante, en la línea de lo utilizado por RSTP en el estándar IEEE 802.1D.

Tramas de Datos

El encapsulado de las tramas de datos en el nivel superior es similar al del nivel superior. Igualmente, un campo Ethernet type permite diferenciar las tramas encapsuladas de las no encapsuladas.

Debido al encapsulado adicional propuesto, la longitud final de una trama de datos es posible que exceda ligeramente la MTU máxima de la LAN. Sobre la longitud de la trama Ethernet IEEE de 1500 octetos (1518 total incluyendo la cabecera de nivel 2 y Frame Check Sequence (FCS)), se denominan *baby giant frames* las tramas que exceden ligeramente la longitud mencionada, mientras que las *jumboframes* son muy superiores orientadas a aumentar el rendimiento. El comportamiento de los bridges comerciales varía de unos equipos a otros, soportando algunos tramas no solamente de las llamadas *baby giant* sino incluso las denominadas *jumboframes* de hasta 9000 octetos [Giant]. La aceptación de tramas superiores al tamaño estándar en ocasiones suele poderse habilitar a nivel de puerto. El problema afecta a todo encapsulado adicional. Una cuestión similar con la longitud de trama ocurrió con el estándar 802.1Q para VLANs, que alargó la trama Ethernet 4 octetos, modificado en el estándar 802.1ac, y en general sucede con cualquier etiqueta adicional Ethernet. Actualmente el encapsulado Q-in-Q supone una extensión adicional, de 4+4 bytes. Una propuesta de manejo de tramas Ethernet de tamaño extendido se expone en [Kap+01]. La solución extrema consistiría en que los bridges ARB y AB fragmenten antes de encapsular una trama que supere la MTU de salida.

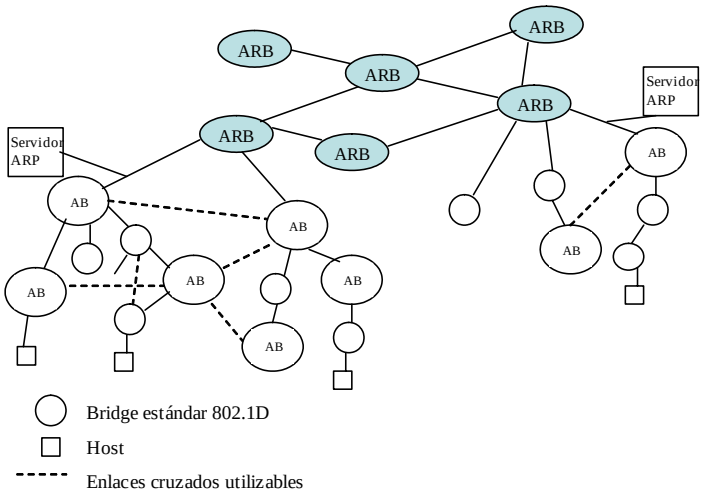


Figura 39. Topología general de red campus.

4.4.6 Servidores

En nuestra propuesta el papel específico más destacado de los servidores se ha descrito para la implementación de ARP con servidores en reparto de carga. Tanto ARP como otros protocolos pueden implementarse de diferentes formas en la red propuesta En la descripción de la propuesta para multicast se detallan los diferentes servicios multicast y las propuestas específicas para ARP, DHCP y otros servicios multicast. Los servidores pueden ser independientes, uno por servicio, o agruparse en una función que llamamos M-server., que abarca diversos servicios que utilizan multicast. Estos servidores tienen como características más destacables el que se descubren entre sí, procesan el paquete según el servicio solicitado y pueden funcionar en reparto de carga. En caso de no poder resolver el servicio, escalan la petición hacia otros servidores o hacia los sistemas finales.

4.4.6.1 ARP

Como se ha descrito en el escenario general de encaminamiento, los servidores ARP resuelven el ARP, recibiendo la trama doblemente encapsulado enviada desde el AB padre, y respondiendo con la MAC del sistema final destino, la MAC del AB destino y la coordenada RSJ más cercana al AB, por si se puede utilizar en el encaminamiento alternativo al Árbol de Expansión si el camino es más corto que el de AE. Los servidores ARP resuelven las direcciones IP de los sistemas finales del area bajo un ARB (Árbol de Expansión completo del área inferior o zona *stub*) y contienen tambien las direcciones MAC y RSJ del AB designado de cada sistema final. En el escenario simplificado, por no existir ABs, el servidor resuelve el ARB destino en vez del AB destino.

Capítulo 5

Protocolos propuestos

En este capítulo se describen los protocolos, nuevos o extendidos, a emplear en las redes campus propuestas. Comenzamos con el nuevo protocolo propuesto para encaminamiento en el troncal AMSTP y el protocolo RSJ, extensión del protocolo STAR, opcionalmente usado en el nivel inferior de la red, así como el nuevo mecanismo RSTAA para asignación de coordenadas en el que se basa. A continuación se describe el nuevo protocolo para reparto de carga entre los servidores/registradores ARP propuestos. El protocolo estándar RSTP se ha descrito en el estado del arte. Se describen las tablas y estructuras de datos utilizadas por los protocolos propuestos, se describe y justifica la propuesta para tratamiento del tráfico multicast y se revisan de forma abreviada, por no ser objetivo específico de la Tesis, las vulnerabilidades más destacables de capa 2 y las medidas de seguridad pertinentes.

5.1 Protocolo Alternativo de Árboles Múltiples de Expansión (AMSTP)

Se describe AMSTP primeramente de forma sumaria y a continuación se describen el plano de control y el plano de usuario. En el plano de control se trata la construcción y mantenimiento de los árboles y en el de usuario el procesado y reenvío de tramas en los bridges. Finalmente se trata el plano de gestión, centrado exclusivamente en los aspectos de configuración de los bridges. Otros aspectos de gestión quedan como trabajo futuro.

Algunas de las actuales posibilidades a este respecto en el plano de gestión se han descrito al revisar el estado del arte [Sha+04][Ish04+].

5.1.1 Sumario

El protocolo AMSTP, propuesto para troncales de redes metropolitanas en [IGA04] y aquí para redes campus, es una evolución de RSTP y MSTP, pudiendo considerarse una variante de MSTP para bridging de caminos mínimos, optimizada para su uso en troncal, mínima configuración y orientada a obtener la máxima eficiencia de uso de la infraestructura existente. El protocolo estándar de árboles múltiples MSTP es muy complejo de configurar y por tanto reduce la fiabilidad de la red, por lo que, entre otras razones, no es muy adecuado para su aplicación en troncales ni en el nivel de Distribución de las redes campus. Las propuestas iniciales en el IEEE [Sea05] relativas a Shortest Path Bridging contienen algunos principios comunes con los expuestos en lo que respecta a la obtención de caminos mínimos utilizando toda la infraestructura propugnando una simplificación de MSTP y diferencias significativas en el resto.

AMSTP es un protocolo de árboles múltiples de expansión que utiliza instancias de árbol originadas en cada Bridge Frontera del troncal para el reenvío de tramas a los bridge destino del troncal (Bridge Frontera destino). Se define como *multiárbol completo* al conjunto de instancias de árbol de distribución, enraizadas cada una en un Bridge Frontera, que interconecta todos los bridges del troncal o núcleo (Core). Los bridges de troncal (*Backbone Bridges*) son bridges ethernet interconectados que ejecutan el protocolo AMSTP entre ellos. En la arquitectura propuesta se denominan Alternative Routing Bridges (ARBs). AMSTP solamente es ejecutado en los puertos del bridge que están conectados a otros bridges del troncal. El resto de los puertos del bridge de troncal ejecuta protocolo RSTP o STP.

5.1.2 Introducción

El protocolo que opera en el troncal de las redes campus avanzadas objeto de esta Tesis es el que denominamos Protocolo Alternativo de Árboles de Expansión Múltiple (Alternative Multiple Spanning Tree Protocol AMSTP). Se utilizan árboles múltiples de expansión de forma simultánea y automática para realizar un encaminamiento eficiente y con distribución de la carga en el troncal. AMSTP está orientado a su compatibilidad con RSTP, protocolo en el que se apoya. Por su simplicidad hace posible el alto rendimiento y la autoconfiguración. Aunque el funcionamiento de AMSTP como protocolo de troncal varía ligeramente según el protocolo utilizado en el nivel de Acceso, se expone de forma separada para su mejor comprensión, las diferencias se muestran al describir los escenarios de encaminamiento simplificado y completo.

5.1.3 Plano de control

5.1.3.1 Construcción de los árboles

El proceso de construcción del multiárbol puede ser descompuesto en dos: construcción del árbol principal y construcción de el resto de las instancias llamadas

Alternate Multiple Spanning Tree Instances (AMSTI) hasta que se construye un árbol por instancia de bridge frontera. Como ejemplo, partiendo de la red troncal que se muestra en la Figura 41, se construirán los árboles que se muestran en la Figura 42 . La construcción del árbol principal funciona de idéntica forma a RSTP. Cada bridge emite autónomamente unidades de datos de protocolo BPDUs cada intervalo de tiempo de saludo (Hello Time) a sus bridges vecinos. En primer lugar el bridge con menor valor de Identidad de Bridge (la cual se forma uniendo la menor prioridad configurada (16 bit) y la menor dirección MAC (48 bit)) es elegido como Bridge Raíz del Árbol de Distribución Principal, ya que cada bridge que recibe BPDUs de dicho bridge, lo adoptará como Raíz y lo anunciará como raíz en las BPDUs que propaga periódicamente. Estas BPDUs (Figura 40), cuyo formato se describe más abajo, contienen el coste del camino desde el bridge que la emite hasta el bridge raíz. Cada bridge se une al árbol de expansión eligiendo como puerto raíz el puerto que recibe la BPDU superior, entendiendo como superior aquella que anuncia el menor coste de camino al bridge raíz. Cada bridge prepara sus propias BPDUs a partir de los datos de las BPDUs recibidas, eligiendo la superiores de acuerdo con los criterios del estándar STP (802.1D): menor identidad de bridge, menor coste de camino, mejor (menor valor de) prioridad de puerto, menor identidad de puerto) y las transmite a través del árbol de distribución principal para el continuo mantenimiento del árbol de distribución óptimo.

La construcción del resto de instancias de árbol, procede de la forma siguiente: cada Bridge Frontera añade a su BPDU básica, elaborada como se ha descrito en el párrafo anterior, la información de todas las instancias AMSTP (árboles) en las que el bridge participa, una por Bridge Frontera. La información que se adjunta por cada instancia de árbol se llama *registro_AM* y contiene información semejante a la usada en la construcción de las BPDU del árbol principal, aplicada a cada instancia, es decir identidad del bridge raíz de dicha instancia, coste del camino desde el bridge hasta el raíz de la instancia e indicadores del protocolo (*flags*). La diferencia fundamental con otros protocolos de árbol de distribución es que en las instancias de árbol que no son la básica no existe proceso de elección de bridge raíz. En AMSTP cada Bridge Frontera se reivindica a sí mismo como raíz del árbol y acepta igualmente las reivindicaciones de otros como bridge raíz de otras instancias de árbol. Dicho árbol es identificado por la dirección origen MAC del bridge raíz. El resto del proceso es análogo a la construcción de los árboles MSTP dentro de una región MST [802.1s]: se eligen caminos de árbol en cada bridge para unirse a el bridge raíz con el mismo criterio de mínimo coste usado en RSTP y MSTP, utilizando los valores de prioridad de puerto e identidad de puerto para deshacer los empates de prioridad. Se utiliza un octeto de indicadores (*flag octet*) por los bridges para comunicarse y negociar transiciones de los estados de puerto y cambios de rol de puerto dentro de cada instancia de árbol.

5.1.3.2 Formato de BPDU

La BPDU del protocolo AMSTP tiene una estructura que se muestra en la Figura 40. Es una extensión de la BPDU del protocolo RSTP [RSTP]. La BPDU básica, compatible con la de RSTP (Figura 3), se utiliza para la negociación y establecimiento del árbol de distribución básico (instancia 0) entre los bridges del troncal. Cada uno de los *registros_AM* anexos se utiliza para negociar y construir una instancia de árbol determinada (AMSTI). El mecanismo de transporte de BPDUs de AMSTP es conjunto, transmitiendo cada bridge la BPDU básica junto con todas los *registros_AM* en una única BPDU para reducir el tráfico de difusión y el procesado de BPDUs en los bridges.

En la propuesta descrita se busca la mayor compatibilidad con el estándar RSTP. Otros formatos de BPDUs no orientados a compatibilidad con RSTP son posibles.

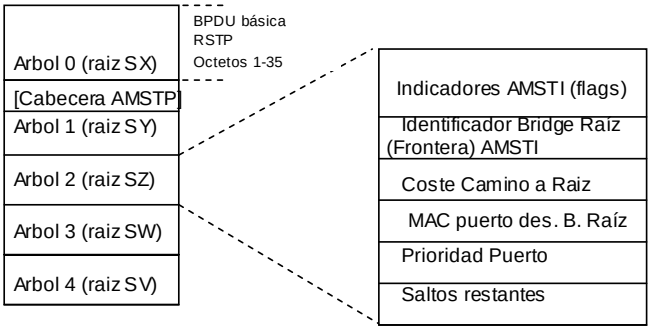


Figura 40. Formato de BPDU AMSTP.

La BPDU AMSTP se compone de la BPDU básica, seguida de la cabecera AMSTP y los bloques *registro_AM* de las demás instancias de árbol. La BPDU básica es compatible con RSTP de forma que un bridge RSTP la interpreta correctamente. Cada *registro_AM* incluye un octeto de indicadores (flags) idéntico al descrito para el árbol RSTP (Figura 3). Estos indicadores se usan por los bridges vecinos en el árbol para negociar las transiciones de una instancia de árbol. Tanto el formato como la utilización de los flags de los *registro_AM* son idénticos a los utilizados para construir el árbol RSTP como se describe en el capítulo 2. Los campos para cada *registro_AM* son los siguientes:

- Indicadores (flags) AMSTP: *cambio de topología, asentimiento de cambio de topología, propuesta, aceptación, estado learning, estado forwarding, rol de puerto: desconocido, alternate, backup, raíz, designado.* Tiene el mismo significado que en RSTP.
- *Identificador Bridge Raíz* (Frontera) AMSTI
- *Coste Camino a Raíz*
- *Prioridad Puerto*: prioridad del puerto para ser seleccionado como raíz
- *Saltos restantes:* para limitar el número de saltos de la BPDU

Los campos de la cabecera AMSTP se utilizan en caso de requerirse compatibilidad con MSTP y/o funcionamiento en modo multi-región. Son los siguientes:

- *Primer Campo*: a cero para compatibilidad con STP longitud versión 0.
- *Long. Versión 4* longitud de la BPDU AMSTP total en octetos.
- *Identificador de Configuración MST.*
- *Identidad de Bridge:* No utilizado.
- *Coste a Bridge Raíz:* No utilizado.
- *Prioridad de Puerto:* Prioridad de elección del puerto.
- *Saltos Restantes:* Saltos restantes de la BPDU.

El mecanismo basado en decremento de saltos restantes limita la difusión y validez de las BPDUs de forma más precisa que el utilizado en STP y RSTP, basado en antigüedad de las BPDUs (Message Age y Max Age).

Al igual que en el protocolo estándar [802.1D], las BPDUS AMSTP se transmiten dentro de una trama LLC de tipo 1 a continuación de los campos DSAP, LSAP y UI. El valor de DSAP y LSAP es el mismo (0x42) que para dichos protocolos.

Se eligen valores numéricos inferiores para indicar información mejor (mayor prioridad) de forma deliberada, tanto para la comparación de coste de camino al bridge raíz, como de prioridades de puerto, para permitir la concatenación de octetos que permite una comparación de vectores de prioridad sin tener en cuenta las longitudes relativas de los vectores comparados.

En los algoritmos de árbol descritos se comparan tanto valores numéricos de identidades de bridge y puerto como prioridades, codificadas como mayor prioridad las de menor valor numérico. Para evitar confusiones se emplea la terminología "*mayor que*", "*igual que*", "*menor que*" cuando se trata de valores numéricos. Para comparar prioridades se utiliza: "*mejor que*", "*la misma que*", "*diferente a*" o "*peor que*". Se utilizan los términos "*superior*" e "*inferior*" para las comparaciones que no incluyen solamente prioridad, sino comparaciones de vectores, por ejemplo el caso que un vector de prioridad más reciente que sustituye a uno más antiguo.

5.1.3.3 Procesado de BPDUs AMSTP. Proceso de instancias múltiples de árbol de expansión

El procesado de instancias múltiples del árbol de expansión comprende dos aspectos: Creación y mantenimiento de la instancia (árbol) propia de árbol de expansión y mantenimiento de las demás instancias. Las instancias de árbol se mantienen mediante las BPDUs emitidas periódicamente por cada ARB de forma autónoma, la cual contiene todos los registros_AM de las instancias de árbol activas. Cada ARB se autoerige en raíz de una instancia de árbol de expansión, para ello inserta un *registro_AM* conteniendo su bridge ID como raíz de la instancia, su identidad como bridge emisor y distancia cero.

El mantenimiento de las demás instancias de árbol requiere que el ARB procese cada *registro_AM* recibido reevaluando las distancias a cada ARB raíz y eligiendo como puertos raíz de esas instancias a aquellos por los que se reciban BPDUs con mejor coste para dicha instancia. La elección de bridge raíz no se produce dado que todos los bridges frontera crean su árbol de expansión. Los bridges de tránsito, es decir los bridges en los que ningún puerto del bridge (ARB) es puerto ARB local (pudiendo ser puertos de tipo 802.1D estándar) no crean su instancia de árbol insertando la información en las BPDUs transmitidas, limitándose a la optimización y actualización de las recibidas eligiendo los *vectores de prioridad* mejores.

Se denomina v*ector de prioridad* a la concatenación de valores que especifican la prioridad de un bridge, puerto o mensaje y permiten compararlo con otro en los algoritmos de árboles de expansión.

El proceso de BPDUs se realiza a nivel de puerto. Llamamos *vector de prioridad de puerto* al vector de prioridad que resulta una vez que la información de las BPDUs recibidas ha sido procesada y cualquier actualización de información ha sido realizada.

El seudo código para el procesado de BPDUs se describe a continuación. La lógica del procesado de árboles de expansión se realiza mediante los llamados *bridge priority vector* (vectores de prioridad de bridge) y los *message priority vector* mensajes de prioridad de bridge, de forma semejante a los estándares [802.1s].

Se define como vector de prioridad de puerto, a la concatenación de valores
vector de prioridad de puerto =
{IDRaíz:CosteInternoRaíz:IDBridgeDesignado:IDPuertoDesignado:IDPuertoReceptor }

Por otro lado se denomina *vector de prioridad de mensaje* al vector de prioridad contenido en un Mensaje de Configuración recibido.

Dado un bridge cuya IDBridge es *J* y recibe por un puerto P_J un Mensaje de Configuración de un bridge D enviado por un Puerto Designado P_D anunciando un Coste Interno hasta Raíz CIR_D, el vector de prioridad es:

vector de prioridad de mensaje = $\{CIR_D: D:P_D:P_B\}$

En AMSTP la comparación de vectores de prioridad de mensaje se realiza de la forma siguiente: Un vector de prioridad de mensaje recibido en un puerto de un bridge AMSTP (ARB) es mejor que el vector de prioridad de puerto si se cumple que,

(CIR < CosteInternoRaíz) || ((D < IDBridgeDesignado)&&(CIR == CosteInternoRaiz))
||
((CIR == CosteInternoRaíz)&&(D == IDBridgeDesignado)&&(P_D *<*
IDPuertoDesignado))

y es igual al vector de prioridad de el puerto si se cumple:

((CIR == CosteInternoRaíz)&&(D==IDBridgeDesignado)&&(P_D *==*
IDPuertoDesignado))

donde *CIR* es el Coste Interno hasta el bridge Raíz de la instancia.

Un vector AMSTI de mensaje de prioridad es superior al vector de prioridad de puerto solamente si el vector de prioridad del mensaje es mejor que el vector de prioridad de puerto, o el Identificador de Bridge Designado y el Identificador de Puerto Designado son iguales en cuyo caso el mensaje ha sido transmitido desde el mismo Puerto Designado como un mensaje superior recibido previamente, es decir:

$\{CIR_D: D:P_D:P_B\}$ es *mejor que*
{IDRaíz:CosteInternoRaíz:IDBridgeDesignado:IDPuertoDesignado:IDPuertoReceptor}
|| $((D==IDBridgeDesignado) \&\& (P_D==IDPuertoDesignado))$

Si el vector de prioridad del mensaje recibido en un Mensaje de Configuración por un puerto Designado es *superior* al vector de prioridad de el puerto, lo reemplazará. Un vector de prioridad de mensaje AMSTI recibido de un bridge que no pertenezca a la misma región es descartado. La *IDRaíz* de la instancia no es necesario compararla ya que es la misma al basarse el árbol en un bridge raíz indiscutido.

5.1.3.4 Difusión de listas de ABs entre ARBs

Los ARBs difunden periódicamente mediante *broadcast* por el árbol de expansión principal que los une (RSTP) la lista de ABs de cada ARB. Cada ARB mantiene una lista de ABs conectados a él. Cada ARB aprende los ABs activos (entre otros mecanismos posibles) porque los ABs envían periódicamente una trama hacia el bridge raíz (ARB) conteniendo su identidad y coordenada RSJ.

Cada *Hellotime* el ARB difunde por el troncal a la dirección multicast *todos los bridges ARB* la lista de ABs. Esta lista es corta por lo que no se consume ancho de banda ni recursos apreciables en los ARBs para su proceso. El número de ARBs es igualmente bajo, por lo que la escalabilidad está garantizada por la jerarquización entre ABs y ARBs. En la sección de Dimensionado se describen los rangos de valores considerados para las redes objetivo. Como mecanismo alternativo a la difusión de listas de ABs entre los ARBs es interesante el aprendizaje simultáneo de parejas de direcciones MAC de ARB y dirección de host cuando el número de sistemas finales no es excesivamente elevado.

La información de estado del protocolo AMSTP se describe, por razones de claridad, conjuntamente con la de los demás protocolos, de manera agrupada por tipo de dispositivo, en la sección Tablas y Estructuras de Datos de este capítulo.

5.1.4 Plano de usuario

5.1.4.1 Procesado de Tramas en Bridges Frontera

En la Figura 41 se muestra el procesado de tramas en el troncal AMSTP. Al procesar una trama, un Bridge Frontera (BF) (*Edge Bridge*) puede actuar como conmutador de entrada, salida o de tránsito. En los dos primeros casos se trata de un Bridge Frontera situado en el límite del troncal. Consideremos primero el caso de un bridge Frontera de entrada. Para procesar una trama unicast procedente del área inferior, el bridge encapsula la trama mediante una cabecera adicional de capa 2 que contiene la dirección MAC del bridge como origen y como dirección destino la dirección MAC del Bridge Frontera destino del troncal. El bridge frontera de entrada SV envía la trama encapsulada por la rama SV-SX-SY de la instancia de árbol de expansión cuyo bridge raíz es el bridge frontera SY de destino.

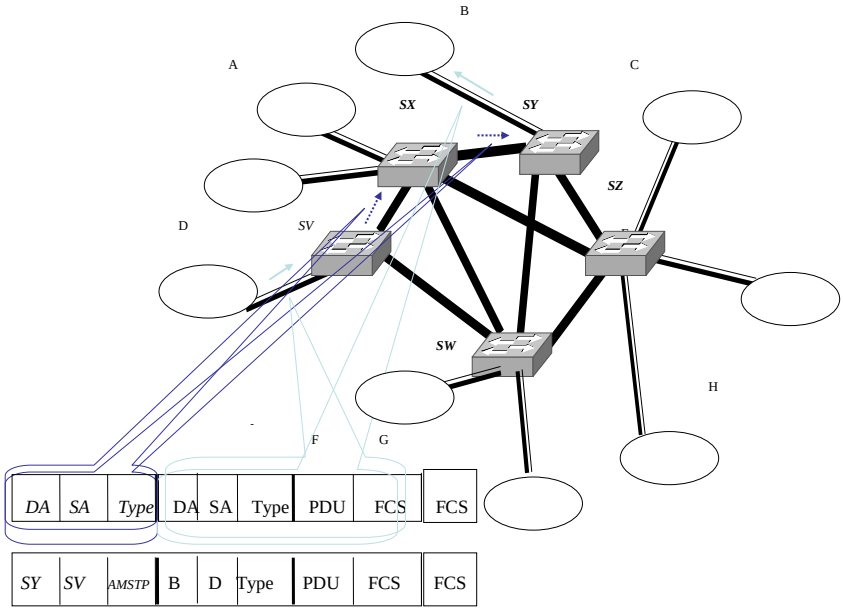


Figura 41. Procesado de tramas en troncal.

Una importante propiedad del camino desde el bridge de entrada al troncal es que el camino hasta cualquier bridge del troncal es mínimo porque el árbol de expansión que se utiliza ha sido construido de forma distribuida, eligiendo cada bridge como puerto raíz el puerto de menor coste hasta el bridge raíz. En la Figura 42 se muestran todos los árboles construidos para la red ejemplo de la figura.

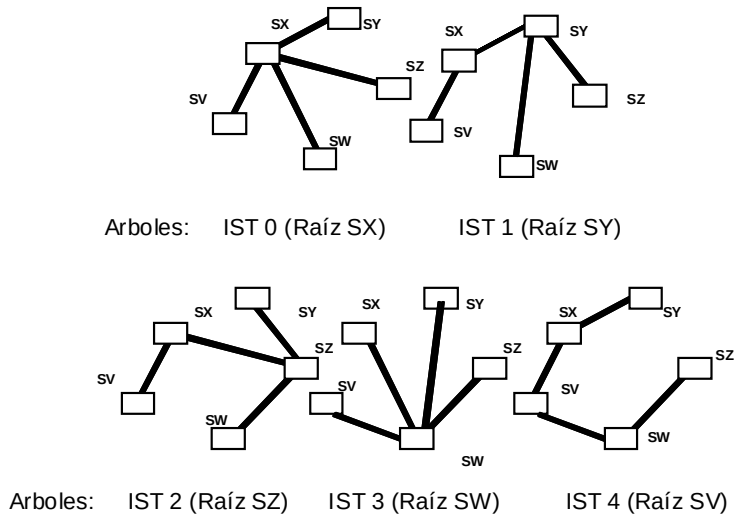


Figura 42. Las cinco instancias del árbol múltiple de expansión alternativo (AMSTP) construidas en la red de la Figura 41.

Los puertos de los conmutadores AMSTP que están conectados a bridges que no ejecutan el protocolo AMSTP no ejecutan dicho protocolo, por lo que son excluidos del mecanismo de reenvío del troncal (de los árboles múltiples de expansión). Con el fin de que los bridges AMSTP de troncal puedan operar con los bridges existentes que ejecutan protocolos como STP y RSTP, parece razonable utilizar un mecanismo como el es protocolo estándar de migración utilizado en MSTP. Básicamente el mecanismo consiste en que si un puerto MSTP escucha BPDUs de versión 0 del protocolo (STP) emitirá solamente BPDUs STP.

5.1.4.2 Procesado en bridges de tránsito

En los bridges AMSTP del troncal que no son bridges frontera, las tramas encapsuladas AMSTP son encaminadas con arreglo a la dirección MAC destino del encapsulado exterior de la trama. Los bridges AMSTP con función pura de tránsito no encapsulan las tramas ni realizan intercambio de listas de AB entre ellos. Establecen las instancias de árboles (AMSTIs) enraizadas cada una en un bridge frontera y encaminan por ellos sin ser raíz de ninguna instancia de árbol.

5.1.4.3 Encaminamiento de tramas

El encaminamiento de tramas en AMSTP se realiza de la forma descrita. Las tramas con doble encapsulado en capa 2 AMSTP se reenvían utilizando el árbol de expansión cuyo raíz es el bridge AMSTP destino de la trama. La trama ingresa por un puerto designado y sale por el puerto raíz del bridge ARB del árbol correspondiente al ARB destino.

Puerto_de_salida = Puerto_raíz(Instancia_árbol(Raíz_ARBdestino))

El bridge ARB destino se obtiene a partir de la dirección MAC DA destino. La dirección MAC del puerto designado del ARB destino ha sido aprendida por las BPDU recibidas, las cuales retransmiten la dirección MAC del puerto origen de esa rama del bridge en el bridge raíz. Las tramas estándar ethernet se encaminan por el árbol RSTP aplicando el aprendizaje de direcciones MAC estándar para reducir la inundación.

5.1.4.4 Formato de trama AMSTP

En la Figura 43 se muestra el formato de trama utilizado en el troncal con AMSTP. El formato es del tipo de encapsulado MAC in MAC. Incluye una cabecera adicional de capa 2 con las direcciones MAC de los bridges frontera origen y destino en el troncal. Incluye la etiqueta 802.1Q con VLAN de uso opcional y un campo TTL para limitar la difusión y como mecanismo antibucle.

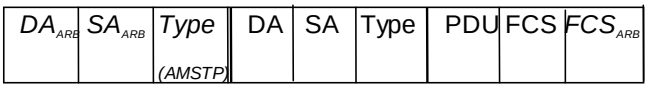


Figura 43. Formato de trama encapsulada (AMSTP).

5.1.5 Procesos AMSTP

Los procesos principales del protocolo AMSTP son los siguientes:

- Mantenimiento del árbol principal.
  - Funcionamiento en modo STP.
  - Funcionamiento en modo RSTP.
- Proceso de instancias múltiples de árbol de expansión.
- Difusión de listas de ABs entre ARBs.
- Encaminamiento de tramas.

5.1.5.1 Mecanismos de compatibilidad con protocolos estándar

AMSTP es compatible con el protocolo RSTP. Por consistencia con los estándares STP y RSTP se recomienda, para la interoperabilidad de AMSTP con RSTP y STP, un mecanismo similar al existente para RSTP para su compatibilidad con el protocolo STP IEEE 802.1D. En la Figura 44 se muestra de forma esquemática el procesado de BPDUs de AMSTP para la elección del modo de operación por puerto según el tipo de BPDUs recibidas.

Para la compatibilidad de AMSTP con STP IEEE 802.1D, un ARB que ejecuta AMSTP detecta en un segmento la ejecución del protocolo estándar STP por el valor a 0 del campo versión en las BPDUs recibidas, en cuyo caso emitirá BPDUs versión 0. El mecanismo establece que cada puerto escucha durante tres segundos las BPDUs recibidas y si detecta una BPDU versión 0 en un puerto, ese puerto pasa a enviar BPDUs versión 0. Si después se detecta una BPDU de tipo 4, pasa a enviar BPDUs versión 4. Lo mismo ocurre con las BPDUs de versión 2 (RSTP) como se indica a continuación.

Para la compatibilidad de AMSTP con RSTP, el ARB detecta la presencia del protocolo RSTP por el valor 2 en el campo versión, por lo que en ese caso empleará BPDUs versión 2. Desde que se detecta una BPDU versión 2 en un puerto del ARB, ese puerto pasa a enviar BPDUs versión 2 quedando en ese estado durante 3 segundos, si después se detecta una BPDU de tipo 4, pasa a enviar BPDUs versión 4. Bajo la implementación propuesta, la identidad de protocolo (protocol ID) a emplear puede ser la misma utilizada por STP (0x42) y diferenciarse de STP y RSTP por el número de versión. Alternativamente puede emplearse una identidad de protocolo específica y única para AMSTP.

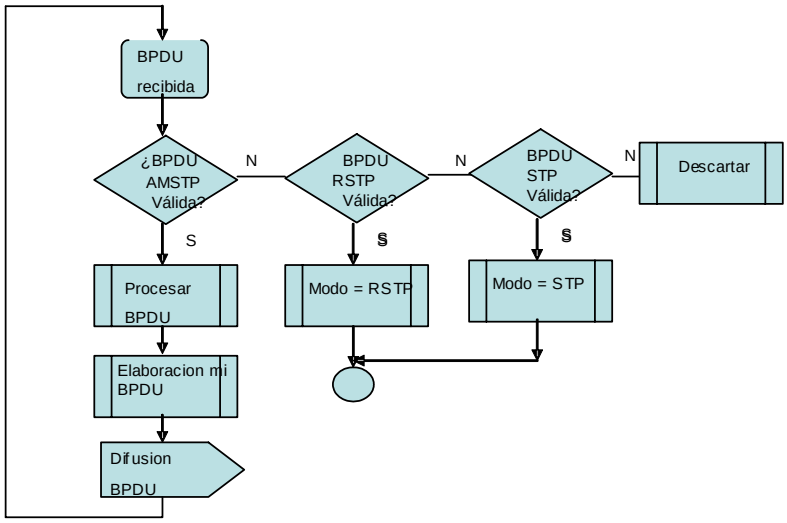


Figura 44. Procesado BPDUs. Migración de protocolo.

5.1.5.2 Procesado de tramas en modo STP o RSTP

Un puerto de un ARB en modo STP o RSTP no procesa las tramas con doble encapsulado ARB (no desencapsula ni encamina) sino que las difunde como cualquier otra trama por el puerto por donde ha escuchado la dirección MAC como MAC origen. En caso de no tenerla en su caché, la difunde por todos los puertos excepto el de entrada.

5.1.5.3 Máquinas de estados.

La implementación de AMSTP se realiza mediante máquinas de estados derivadas de MSTP, interconectadas según se muestra en la Figura 45. Al igual que en MSTP hay máquinas de estado por bridge, por puerto y por árbol y puerto. La máquina de estados de transición de roles de puerto es única por bridge por controlar las transiciones del árbol principal. Las de migración de protocolo, transmisión y recepción de puerto y de cambio de topología son por puerto, las de información de puerto y transición de estados de puerto son por árbol y puerto para mantener los árboles respectivos.

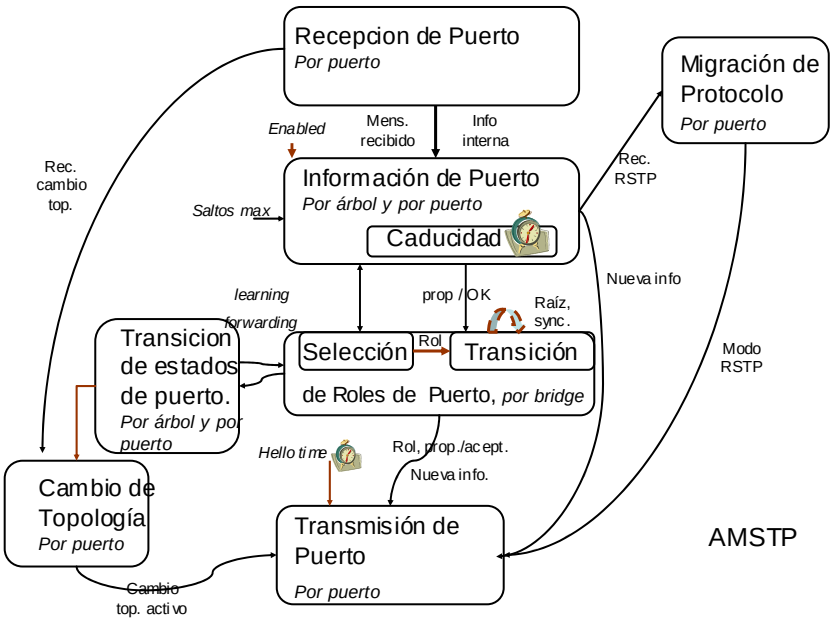


Figura 45. Relación entre máquinas de estados AMSTP.

5.1.6 Plano de gestión

En este epígrafe se tratan los aspectos de gestión y configuración del protocolo AMSTP y se razonan las decisiones de diseño. El análisis se limita a los aspectos de configuración, básicamente relativos a la existencia o no de encaminamiento por VLANs en el núcleo y al número de regiones AMSTP a soportar, algunos de los cuales se han mencionado ya al describir los dispositivos. Como se establecía en la definición del problema, uno de los requisitos básicos es que la necesidad de configuración de los dispositivos sea mínima o nula. Se analiza también la problemática de las regiones AMSTP y alternativas de implementación, según los ámbitos de aplicación del protocolo AMSTP.

5.1.6.1 VLANs y configuración.

Las VLAN son utilizadas actualmente en las redes campus para separar tráfico de difusión creando redes distintas sobre un misma red física. También se emplean para facilitar el uso diversificado de la infraestructura de red mediante árboles múltiples ligados a grupos de VLANs.

En la propuesta de red se distingue entre núcleo y red inferior. Si se usa AMSTP en el núcleo, no procede la separación de tráfico por VLANs, dado que se utiliza un mecanismo paralelo al de VLANs con el fin de utilizar toda la infraestructura, pero no ligado a VLANs. Aunque es posible concebir mecanismos de separación adicionales basados en VLAN para la difusión selectiva en el troncal, se considera que iría en detrimento de la simplicidad sin grandes beneficios de prestaciones.

En el nivel inferior pueden existir las VLAN de la forma estándar, que serán encapsuladas de la forma estándar en el núcleo. Los protocolos de trunking de VLANs y de registro deben operar de la forma estándar en la red campus, de forma que sea posible crear y registrar VLANs en los bridges de las redes del nivel inferior.

La necesidad de mínima configuración es una característica fundamental de los bridges transparentes, atenuada actualmente por la complejidad creciente de los mismos y por el empleo de VLANs, pero se considera un requisito importante para los bridges de troncal. La utilización de VLANs y del protocolo estándar MSTP para configurar árboles de expansión por grupos de VLANs permite una utilización eficiente de la infraestructura de comunicaciones si bien al precio de una gran complejidad de configuración. Esto se debe a la necesidad de diseñar y planificar los árboles de expansión y de configurar la correspondencia entre las VLANs y los árboles mediante una tabla de hasta 4096 entradas. En el caso de MSTP, esta configuración se debe realizar en cada bridge por no existir medios de configurarlo por áreas. La única simplificación posible hoy consiste en que el sistema de gestión asuma la coordinación y verificación de todas las tablas de los bridges de la región MSTP, lo que puede introducir restricciones adicionales de compatibilidad en las interfaces de gestión. Los errores de configuración, altamente probables, se traducirán en fallos de operación. El protocolo AMSTP simplifica el proceso utilizando instancias de árboles creadas automáticamente en lugar de asignarlas a VLANs, de forma que no es necesario este mapeado en los bridges. Los parámetros a configurar son comunes a RSTP, tales como selección del bridge raíz (configurando la prioridad más baja) y configuración de los bridges de reserva de la región y sus prioridades.

5.1.6.2 VLAN por bridge frontera

En el modo de trabajo compatible con MSTP que hemos descrito para los bridges de tránsito, cada bridge frontera crea una VLAN ID que inserta en las tramas en el troncal. De esta forma los puertos de entrada a los bridges frontera son asignados implícitamente a una VLAN a la que pertenecen por razón del *bridge frontera.* Este es un criterio nuevo de pertenencia a VLAN que seguramente debiera ser estandarizado[2]. Es un criterio ligado a condiciones que describen flujos, en otras palabras, al flujo que se origina en el bridge frontera. El tráfico en sentido opuesto, desde los demás bridges frontera a este bridge, se procesa por los árboles de los otros extremos, por estar originado en dichos bridges, y pertenece a cada VLAN correspondiente al bridge frontera de destino. Esta VLAN queda definida por el tráfico insertado al troncal desde el árbol inferior. No es preciso asignar una VLAN explícita a dichos árboles.

5.1.6.3 Regiones AMSTP

Se analiza ahora la problemática de las regiones AMSTP, entendiendo por región el área formada por uno a N bridges y sus enlaces, en las que se establecen los árboles múltiples AMSTP. Se analiza en relación a MSTP y a compatibilidad con MSTP.

En el diseño de AMSTP pueden considerarse tres alternativas: ausencia de regiones, región única y regiones múltiples. En su aplicación de redes campus de mínima configuración, AMSTP es de red única para evitar la complejidad de configuración derivada de la multiplicidad de regiones a definir. En otras aplicaciones, la decisión dependerá de la importancia relativa de aspectos como la simplicidad de configuración

[2] La reciente propuesta de Shortest Path Bridging parece confirmar esto.

respecto a la posible compatibilidad con otros protocolos como MSTP. En los escenarios propuestos la región única se adapta mejor a los requisitos establecidos en el capítulo 2. A continuación se analizan los aspectos de diseño afectados.

Consideremos el caso de ausencia de regiones. Nos referimos a ausencia de regiones si en el protocolo no se contemplan o identifican regiones. Un ejemplo es RSTP. Región única se refiere a que no se contempla la posibilidad de varias regiones separadas AMSTP, siendo solamente una aceptable, pero se incluyen identificadores de región que puedan hacerlo compatible con MSTP, protocolo que sí contempla regiones.

Consideremos el caso de región única. La forma en que en el protocolo MSTP un bridge conoce a qué region pertenece y cómo verifica que está en su región es la siguiente: En MSTP se identifica una región por una tabla de asignación de VLANs a árboles (instancias) asociada a un identificador de tabla y de versión más el resumen del contenido de dicha tabla obtenido mediante *hash* MD5. Esto se transmite en las BPDUs para que cada bridge pueda verificar la coincidencia de identificador de configuración y su integridad y con ello su pertenencia a la región y el correcto mapeado de VLANs a árboles de expansión. Si solamente existe una región AMSTP coincidente con el núcleo (core) o troncal de la red, no habrá regiones AMSTP colindantes y no habrá bridges conectados que ejecuten AMSTP y pertenezcan a otro troncal. Tampoco habrá en principio tabla de mapeado de VLANs a instancias de árbol ni por lo tanto necesidad de verificar la integridad e idéntico contenido de la misma en todos los bridges del área. Lo que es siempre necesario incluso en entornos de región única como los que nos ocupan es que los bridges detecten los bordes de la región. Esto se realiza en los puertos de los bridges frontera en los que se escuchen BPDUS no AMSTP (STP, RSTP) de otros bridges conectados.

Respecto a la utilización de AMSTP en regiones múltiples adyacentes, AMSTP está inicialmente concebido para su aplicación en entornos de región única como los núcleos de redes campus. Sin embargo es aplicable en regiones múltiples. Para ello es preciso configurar en cada bridge un identificador de región que permita a los bridges identificar los bordes de la región. La detección de borde se realiza en cada puerto. Parece factible implementar AMSTP utilizando los mecanismos de MSTP tales como el árbol CIST para enlazar las regiones y un IST interno en cada región para la conectividad dentro de la misma. Frente a MSTP se mantendría la ventaja relativa de configuración, pero la complejidad de las redes resultantes se aproxima a las de MSTP. El análisis de esta configuración está fuera del objetivo de esta Tesis.

En el caso de desear aplicar el protocolo AMSTP en un entorno multirregión en el que existan redes MSTP, la configurabilidad obligaría en ese caso la compatibilidad con MSTP. Para ello el formato de las BPDU AMSTP debería ser compatible con MSTP y la ventaja de simplicidad de configuración de AMSTP prácticamente desaparece como se describe a continuación.

5.1.6.4 Compatibilidad AMSTP-MSTP en núcleo

Es posible, (aunque contrario a los requerimientos de configuración simple), compatibilizar el funcionamiento de bridges frontera ARB con protocolo AMSTP con bridges de tránsito del núcleo con protocolo MSTP. En este epígrafe se describen las implicaciones de configuración.

Si se desea compatibilidad en los bridges AMSTP con el protocolo MSTP se requiere una adecuada configuración. El protocolo IEEE estándar vigente para árboles múltiples es MSTP. Actualmente MSTP admite como máximo hasta 4096 VLANs y hasta 64 instancias de árbol. En MSTP, la VLAN ID se transmite insertada en la trama (etiqueta 802.1q) y mediante dicha VLAN ID se determina en cada bridge el árbol aplicable consultando la tabla de asignación de VLAN a instancias de árbol. Esta tabla se debe configurar en cada bridge de la región de forma idéntica. Adicionalmente, en la identidad de Bridge transmitida en los *registro_AM* para cada árbol, se puede incluir la VLAN (12 bit) en el campo de prioridad del bridge para ahorrar direcciones MAC de bridges a utilizar, empleando una única MAC por bridge para todas las VLAN de las que participa. De esta forma los 4 bits restantes mas los 12 para la VLAN ID forman la identidad del bridge cuando se usa la característica de ahorro de direcciones MAC.

La configuración estándar MSTP requiere establecer una correspondencia de VLANs a árboles. Para compatibilizar AMSTP y MSTP debe asignarse solamente una VLAN a cada árbol. Esta VLAN debe estar asociada a un bridge frontera. Para ello debe elaborarse la lista de bridges frontera y establecer su correspondencia a sus VLAN, o utilizar un protocolo especial de gestion. Los bridges de tránsito siguen a los bridges frontera en el establecimiento de árboles, aceptándolos como bridge raíz (su prioridad deberá ser superior) , establecen las instancias MSTIs de la que ninguno de ellos es raíz y reciben configuración de VLANs que les indican a que MSTI se mapea cada VLAN.

Para la compatibilidad AMSTP-MSTP, la lista de bridges frontera mencionada se instrumenta como en MSTP en forma de tabla de correspondencia entre VLANs e instancias de bridges identificada por una *configuration_ID*. Esta *configuration_ID* debe ser emitida por los bridge frontera en las BPDU con el fin de que los bridges MSTP verifiquen la pertenencia a dicha región MSTP. Si se establece adecuadamente la configuración en los bridges de tránsito de manera que actúen coordinadamente con los bridges AMSTP pueden funcionar de forma compatible. En este caso se requiere que los bridges AMSTP emitan las tramas etiquetadas con una identificación de VLAN asignada previamente de forma que los bridges de tránsito MSTP puedan asignar cada trama al árbol correspondiente. Adicionalmente, los bridges frontera deben emitir en los *registro_AM* su identidad con una prioridad mínima para asegurar su elección como bridges raíz de su árbol por los bridges de tránsito. Las BPDUs AMSTP deberán pues ser totalmente compatibles en formato con las BPDU MSTP, para ser procesadas adecuadamente por los bridges de tránsito MSTP.

5.1.6.5 Doble aprendizaje MAC en entornos de troncal MAN

AMSTP es aplicable en diferentes escenarios. En el escenario propuesto en esta Tesis el número de equipos terminales es muy alto y se utilizan mecanismos especificos para el encaminamiento apoyados jerárquicamente en los ABs. En los escenarios de redes troncales de Redes de Area Metropolitana como se muestra en la figura el número de MACs a aprender es bajo porque el dominio conmutado termina en los encaminadores. En este caso los bridges de troncal aprenden las direcciones MAC de otros conmutadores del troncal de las tramas recibidas con doble encapsulado. Asimismo aprenden las direcciones MAC de los nodos terminales conectados al troncal (nivel inferior de la red) inspeccionando las direcciona MAC internas. Denominamos a este proceso de aprendizaje MAC como *aprendizaje doble MAC.* En cada puerto del bridge se construye una tabla que contiene los pares de direcciones MAC aprendidas en ese puerto por la

inspección de la cabecera de las tramas recibidas en donde la dirección de la cabecera de capa 2 interior identifica el destino remoto a incluir en la tabla y la dirección externa MAC origen se utiliza para seleccionar la tabla de reenvío (árbol) con la que se debe trabajar. Cuando el Bridge Frontera tiene que reenviar una trama, realiza una búsqueda en las tablas de la dirección MAC destino del paquete y obtiene la MAC del bridge destino así como el puerto del bridge por el que el troncal bridge destino ha sido aprendido, es decir, el puerto por el que se recibieron las tramas que lo contenían en su cabecera. Si no hay información en la tabla de la dirección MAC destino en la tabla la trama se envía encapsulada con la dirección MAC multicast reservada para "todos los bridges de protocolo AMSTP"

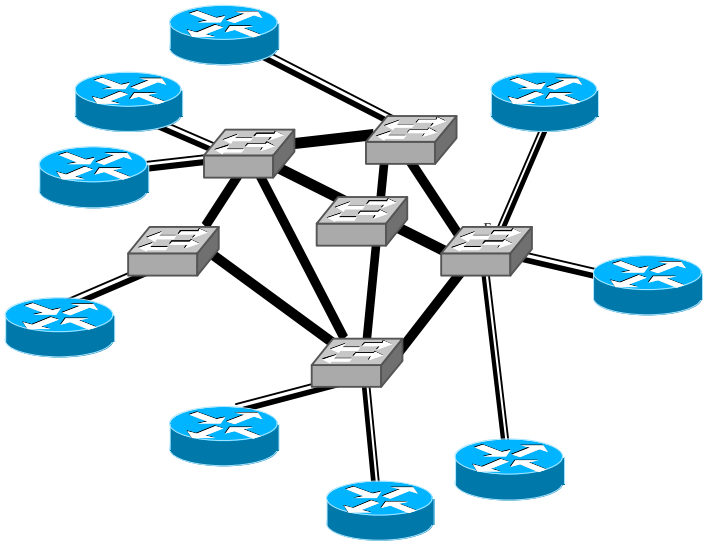


Figura 46. Red troncal con encaminadores.

5.1.7 Resumen de características AMSTP

Se revisan ahora las características principales del protocolo AMSTP. AMSTP está enfocado al núcleo por lo que su diseño busca la simplicidad que permita el máximo rendimiento en la conmutación.

En la Tabla 3 se muestran las principales características del protocolo AMSTP. La primera de la tabla se refiere al criterio que se aplica en el bridge para asignar una trama a un árbol u otro de los muchos posibles. AMSTP utiliza la dirección MAC origen del encapsulado exterior de capa 2 (cabecera MAC añadida por el bridge Frontera de entrada). La segunda se característica se refiere a la construcción de las instancias de árbol, cada instancia se forma automáticamente asociada a la bridge_id de un bridge frontera del troncal. La tercera es el número de instancias de árbol, en AMSTP se crea una instancia de árbol por cada bridge frontera por ser cada uno raíz de una instancia. La siguiente se refiere al procedimiento de elección de bridge raíz de cada instancia, radicalmente simplificado respecto a otros protocolos de árboles de expansión, a fin de

garantizar que los árboles se construyen partiendo de las fuentes de tráfico, es decir, de los bridges frontera. Esta característica facilita al mismo tiempo la autoconfiguración (evitando la planificación detallada de VLANs y su asignación a árboles de expansión) y la obtención de caminos mínimos para tráfico unicast y multicast.

Tabla 3. Principales características AMSTP

Característica del protocolo	AMSTP
Criterio de pertenencia de una trama a un árbol (instancia)	Dirección MAC origen de la cabecera exterior (encapsulado L2)
Criterio de formación del árbol	Por dirección MAC del bridge frontera
Número de árboles	Uno por bridge frontera
Preconfiguración de árboles	Automática
Elección de bridge raíz por árbol	Sin elección. Cada bridge es raíz de u instancia de árbol..
Bridge ID	Estrictamente sólo se requiere la dirección MAC.
Número de regiones	Unica

5.1.8 Árboles independientes (N*RSTP).

Bajo la denominación N*RSTP identificamos un concepto, aún pendiente de análisis detallado que el autor considera pertinente dentro del espacio de soluciones basadas en árboles múltiples de expansión de caminos mínimos. Las ideas básicas se exponen a continuación y son resultado de la proyección de las ideas expuestas anteriormente. Se parte de un número ***N*** de instancias de árboles de expansión igual al número de nodos de la región. Cada nodo emite BPDUs de un protocolo de árbol de expansión de tipo RSTP de forma independiente como raíz de un árbol de expansión. Estas BPDUs se distinguen entre sí por la bridge ID contenida en el campo bridge raíz, origen de dichas BPDUs y por una nueva variante de la identidad de protocolo. Cada bridge emite BPDUs basadas en la información de las BPDUs recibidas, de forma similar a *n* árboles de expansión RSTP que operen independientemente. En el modo compatible RSTP, los bridges utilizan un solo árbol RSTP, el establecido de forma convencional, con elección de bridge raíz. Cada bridge encamina las tramas al bridge destino por el árbol basado en el bridge destino. Para la difusión de las tramas multicast y broadcast se utiliza el árbol del bridge origen. Para las tramas de vuelta, se usa el árbol del bridge origen (destino de la trama de vuelta). Si no está disponible la instancia del árbol de destino para encaminar, se debe descartar la trama, porque aunque podría enviarse por el árbol del bridge origen existen peligros de bucle si unos bridges encaminan por árbol origen y otros por destino. La diferencia básica consiste en la utilización de tantos árboles independientes de expansión como nodos activos existan en el núcleo mediante ejecuciones completamente independientes del protocolo RSTP simplificado.

La región *multi rstp,* al igual que ocurre con MSTP, debe ser delimitada, así como su posible compatibilidad. La integración con un entorno multi región como el de MSTP requiere un único IST en la región y unirse a un CST global de todas las regiones. En los

entornos MAN y campus el uso de encapsulado puede servir para aislar regiones sin los inconvenientes de complejidad de configuración de MSTP (que delimita regiones mediante configuración) si bien reduciendo la compatibilidad, dado que los bridges que encapsulan introducen una discontinuidad respecto a los bridges MSTP.

Al igual que sucede en AMSTP y GOE, si no se precisa aprendizaje de direcciones MAC ajenas a las de los ARBs, no es necesaria la coincidencia de árboles en ambos sentidos descrita en Shortest Path Bridging para el adecuado aprendizaje de direcciones MAC y coincidencia de camino de la pareja de VLANs de ambas direcciones. Se encaminan las tramas a través del puerto raíz elegido en la instancia de árbol correspondiente, de cada bridge del camino, hacia el bridge destino (raíz del árbol).

5.2 Asignación Automática de Coordenadas de Árbol de Expansión (RSTAA).

En esta sección se describe brevemente el mecanismo o protocolo de asignación de coordenadas denominado RSTAA (Rapid Spanning Tree Based Address Asignment Protocol). En lo sucesivo se denomina como protocolo, pero normalmente se implementará como mecanismo formando parte de otro protocolo, como RSTP.

Este protocolo se utiliza para la asignación automática de las coordenadas jerárquicas empleadas por el protocolo RSJ, utilizado en la capa inferior de la arquitectura propuesta como encaminamiento opcional complementario al de árbol de expansión. Se proponen algunas posibles aplicaciones adicionales a la ya mencionada del protocolo RSJ.

5.2.1 Protocolo RSTAA

Tanto el direccionamiento como el encaminamiento jerárquicos son reconocidos por su efectividad y escalabilidad. Sin embargo presentan los inconvenientes de requerir configuración adicional y protocolos para crear y mantener la jerarquía de nodos. La diferenciación funcional creada por la jerarquía en los nodos es también un inconveniente adicional. Dentro de la complejidad creciente de los bridges Ethernet (VLANs, inspección IGMP, etc), este protocolo se inscribe dentro de la tendencia a la autoconfiguración de redes conmutadas, redes que se desean del mayor tamaño posible, dado el predominio de Ethernet en las redes campus y su creciente importancia en las redes metropolitanas.

Este protocolo de asignación de coordenadas utiliza la jerarquía de nodos creada por el Protocolo Rápido de Árbol de Expansión mediante la identidad de los puertos designados como identificador en cada nivel jerárquico. De esta forma, la coordenada jerárquica que un bridge obtiene automáticamente consiste en la cadena de las identidades de los puertos designados desde el bridge raíz hasta el bridge en cuestión.

Estas coordenadas expresan la posición topológica del bridge en el árbol RSTP. Por razones de compatibilidad, estas coordenadas no pueden emplearse como alternativa a las direcciones MAC ni a las direcciones IP en los protocolos estándar, sino como ayudas al encaminamiento, como identificadores de nodo o en otras aplicaciones como agrupamiento de nodos o elección de líder en redes fijas. El protocolo requiere, al igual que RSTP, que todos los enlaces Ethernet entre bridges sean dedicados. El protocolo es distribuido y simple, suponiendo un incremento muy pequeño sobre el protocolo RSTP.

Expliquemos brevemente el sistema de direccionamiento propuesto basado en la topología de la red. Se apoya en el reciente Protocolo Rápido de Árbol de Expansión (Rapid Spanning Tree Protocol (RSTP), utilizando la información de conectividad relativa entre Puentes, empleando los números de puerto de los puertos designados como coordenadas de cada enlace tomando como origen el Bridge Raíz (Root Bridge) como se muestra en el ejemplo de la Figura 47. El Bridge Raíz, seleccionado de acuerdo con el protocolo RSTP de acuerdo con el valor de su identidad con el prefijo de Prioridad programado, es el origen de coordenadas de la red campus, el punto común de referencia de direcciones. El Bridge Raíz se identifica por su Bridge ID transmitida en las BPDUs del protocolo RSTP. Se asume que la red campus tiene enlaces dedicados punto a punto entre todos los Routing Bridges. En los enlaces entre RBs y sistemas finales pueden existir enlaces compartidos (*leaf nodes* en terminología STP).

La coordenada *rstaa(B)* de un bridge B o host se forma concatenando las identidades de puerto designado, en sentido descendente del árbol de expansión, desde el bridge raíz hasta el puerto raíz del bridge en cuestión:

$$rstaa(B) = id\_puerto\_n0.id\_puerto\_n1.....id\_puerto\_nk$$

donde *id_puert_nj* es la identidad de puerto (de 0 a 1023 según el estándar IEEE 802.1D) del puerto designado que une a cada bridge con su bridge antecesor dentro del árbol de expansión. En el ejemplo de red de la Figura 47 el bridge unido al servidor tiene la coordenada RSTAA 14.570 porque el camino para alcanzarlo a través del árbol de expansión hacia abajo comienza en el Puerto número 14 del Bridge Raíz (Puerto Designado) y pasa por el puerto 570 del siguiente Bridge de inferior nivel del STP. En aplicaciones distintas de RSTAA a la de esta Tesis, si las coordenadas RSTAA se aplicaran a sistemas finales en vez de a bridges, éstas se extenderían un nivel más añadiendo el número del Puerto Designado correspondiente del último bridge, extendiendo el requisito de enlaces dedicados hasta el último tramo de la red. En nuestro ejemplo el enlace al servidor tendria una dirección 14.570.235 siendo 235 el número de puerto del Bridge 14.570 que lo une al servidor.

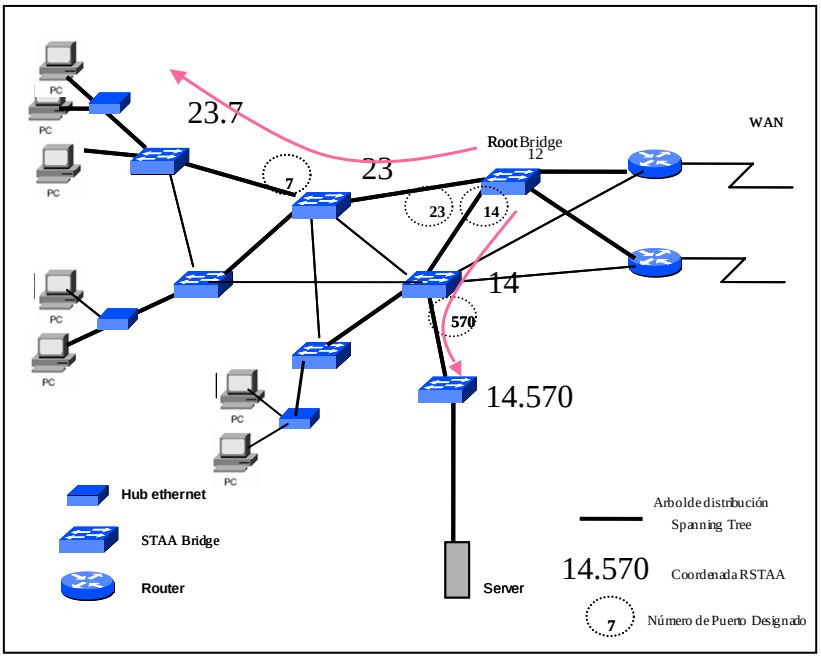


Figura 47. Asignación de coordenadas mediante árbol rápido de expansión (RSTAA).

En la Figura 48 se muestra un ejemplo de BPDU RSTAA. Se utiliza un identificador de protocolo SSAP/DSAP propio y distinto al de STP BPDU (0x42h) para identificarlas. El campo de version identifica la version del protocolo y el tipo de mensaje básico es de tipo *coordenada*. El campo *longitud* indica el número de identidades de puerto encadenadas, el número de niveles de que consta la coordenada RSTAA y se corresponde con la altura relativa del bridge dentro del árbol de expansión. La BPDU RSTAA incluye la identidad del bridge raíz seguida de las identidades de puerto designado que forman la coordenada RSTAA y finalmente la identidad del bridge que emite la BPDU.

ID Protocolo (2 octetos) (RSTAA)
Versión (1 octeto)
Tipo de mensaje (Coord. Bridge)
Longitud (1 octeto)
ID Bridge (8 octetos)
ID Bridge Raíz (8 octetos)
ID Puerto Designado (2 octetos) 1er nivel
ID Puerto Designado (2 octetos) 2º nivel
ID Puerto Designado (2 octetos) 3er nivel
.....

Figura 48. Formato BPDU RSTAA.

5.2.1.1 Borrado y reasignación de coordenadas RSTAA

Aunque a primera vista las coordenadas RSTAA podrían parecer volátiles debido a los cambios de topología en la red campus, ha de tenerse en cuenta la aplicación actual y real de STP en general y las características específicas de RSTP: el Bridge Raíz se elige de forma cuidadosa y prácticamente programada, dado que se configura la prioridad del mismo y del resto para que el Bridge Raíz y sus posibles sustitutos estén previstos y la topología resultante controlada. Por otra parte, la rapidez de reconfiguración de RSTP hace que la actualización de las coordenadas RSTAA sea prácticamente inmediata a dicha reconfiguración.

Al producirse un cambio de topología en el árbol de expansión, el protocolo RSTP distribuye de inmediato una notificación de cambio de topología por todo el árbol en las BPDUs (TCN). Esta notificación produce, además del borrado estándar RSTP de las direcciones MAC aprendidas por los bridges, el borrado inmediato de las coordenadas RSTAA del bridge. Hasta que el protocolo RSTP vuelva a pasar al estado Designado sus puertos no se vuelven a emitir BPDUs RSTAA conteniendo las coordenadas. La emisión de dichas BPDUs está pues vinculada totalmente al estado Designado de los puertos.

5.2.1.2 Operación del protocolo RSTAA

El protocolo RSTAA opera de manera complementaria y combinada con RSTP, dependiente de RSTP pero sin interacción, utilizando solamente información de estado del protocolo para incluirla en las BPDUs RSTAA, siendo éstas independientes de RSTP a fin de preservar la compatibilidad de RSTP. Las BPDUs RSTAA se retransmiten periódicamente, al igual que las RSTP BPDUs y por separado, entre los bridges vecinos. Los bridges emiten BPDUs RSTAA en cada uno de sus puertos designados hacia los bridges Designados. Cada BPDU contiene la coordenada asignada a dicho puerto siguiendo el árbol de expansión de forma descendente, por lo que la BPDU transmitida en cada puerto es ligeramente diferente, difiere en el último número de puerto. Los puertos que no están en estado Designado no transmiten BPDUs RSTAA. El proceso RSTAA opera en cascada y como esclavo de RSTP, no existen decisiones autónomas en el protocolo RSTAA sino que se distribuye información producida por el protocolo RSTP para su aplicación como coordenadas. Con este enfoque, podemos considerar el protocolo RSTAA como una extensión del protocolo RSTP.

El proceso opera como sigue: una vez que el bridge raíz acuerda con sus bridges designados la transición de dichos puertos a estado de reenvío (*forwarding)* mediante el mecanismo de acuerdo rápido local, dicho bridge transmite hacia abajo BPDUs RSTAA conteniendo su coordenada RSTAA.

En el caso del bridge raíz no existen coordenadas previas recibidas de un bridge antecesor por lo que su coordenada es 0 y solamente incluye en la BPDU la identidad de cada puerto designado a través del cual se está enviando dicha BPDU al bridge Designado respectivo. En otras palabras, el bridge raíz asigna una coordenada distinta de primer nivel (entre 0 y 1023) a sus bridges designados.
Estos bridges preparan sus BPDUs RSTAA a su vez, añadiendo la coordenada de segundo nivel (identidad de puerto de segundo nivel) en la BPDU según el puerto

utilizado para transmitirla y así sucesivamente. Cada bridge incluye asimismo su bridge ID y la del bridge raíz.

El mecanismo de acuerdo rápido opera la forma siguiente: los Bridges designados, al pasar su puerto raíz a estado de reenvío por acuerdo con el bridge superior, bloquean sus puertos designados para así impedir los bucles en la red, propagándose la puesta en estado de reenvío del árbol de expansión con un nivel cada vez, en sentido descendente del árbol. Cualquier cambio en el estado de puertos del bridge que implique cambio de coordenadas RSTAA es detectado mediante los indicadores de cambio de topología en las BPDU RSTP recibidas y produce la cancelación inmediata de las coordenadas RSTAA y el cese de su transmisión.

5.2.1.3 Formatos de coordenada RSTAA

Las coordenadas RSTAA son intrínsecamente de longitud variable. Su codificación, puede ser de longitud fija o variable según convenga en el protocolo que las utilice. Dado que el diámetro máximo de red recomendado por IEEE 802.1D es de 7, el formato fijo recomendado es el siguiente: De 64 bits de longitud: 3 bit (longitud coordenada) - 10 bit primera coordenada- 2ª-3ª……- 6ª coordenada. Formato: *nnn-1111111111-2222222222-3333333333-4444444444-5555555555-6666666666.*

5.2.1.4 Pseudocódigo

El seudocódigo para la elaboración de la coordenada RSTAA del bridge, a partir de la BPDU RSTAA recibida en el puerto raíz del bridge es el siguiente:

Do for ID_puerto = 0, Num_puertos
If estado puerto (numero_puerto)= designado
RSTAA_BPDU_a_enviar = Ult_BPDU_RSTAA_recibida
Ult_BPDU_ RSTAA_BPDU_recibida.longitud = Ult_….recibida.longitud + 2
Añadir ID_puerto j = ID_puerto_designado (este bridge)
enviar RSTAA_ BPDU_a_enviar
else
endif
end

Además de las BPDU RSTAA descendentes, los bridges transmiten BPDUs ascendentes, con tipo de mensaje *coordenada_asignada*, que son reenviadas por cada puerto raíz hasta alcanzar el bridge raíz. Estas BPDUs son opcionales y tienen por objeto permitir conocer la profundidad del árbol de expansión en cada rama. En el protocolo RSJ pueden utilizarse para optimizarlo y adecuar sus parámetros a la profundidad actual de la red. Sirven asimismo como mecanismo de verificación y monitorización del protocolo de asignación de coordenadas.

Aunque pueda argumentarse con razón que las direcciones IP y MAC son suficientes (y quizás excesivas por su independencia mutua, al menos en IPv4), existen algunas ventajas destacables de las coordenadas RSTAA. La primera ventaja es que por ser jerárquicas permiten la agregación de rutas para el encaminamiento y por tanto tablas de encaminamiento de menor tamaño (las direcciones MAC no son agregables por ser planas) y contienen información topológica que puede ser empleada por l protocolos de encaminamiento adecuados. La segunda ventaja es el automatismo en la asignación de

coordenadas, evitando la configuración manual y sus errores. Respecto a los inconvenientes, el primero es que, dada la estandarización y difusión de las direcciones IP y MAC, las coordenadas RSTAA quedan excluidas del encaminamiento explícito estándar de nivel 2 y 3 y deben usarse de forma complementaria utilizando encapsulado adicional.

5.2.1.5 Aplicación de RSTAA a sistemas finales.

Se ha descrito hasta aquí la asignación de coordenadas RSTAA a bridges. El principio es aplicable igualmente para asignar coordenadas a los sistemas finales (SF). El requisito indispensable continúa siendo el mismo: que todos los enlaces Ethernet sean dedicados para posibilitar la operación de RSTP y al mismo tiempo la asignación unívoca de coordenadas. Para la operación del protocolo es suficiente que los sistemas finales procesen las BPDUs enviadas por los bridges conteniendo la coordenada RSTAA asignada. Sin embargo la utilización de este tipo de coordenadas por los sistemas finales en procesos de encaminamiento parece poco práctica (salvo en redes especiales no sujetas a requisitos de compatibilidad) dado el impacto requerido en la funcionalidad de los sistemas finales, perdiéndose la muy recomendable transparencia de los bridges. Adicionalmente, el requisito de enlaces dedicados a nivel de sistemas finales es más exigente cara a la compatibilidad con la base instalada de enlaces compartidos y concentradores *(hubs)* Ethernet.

La aplicación de RSTAA a sistemas finales permite asignar una coordenada topológica a cada sistema conectado a la red. Esto puede tener ventajas en escenarios de movilidad. Esta coordenada identifica el punto de conexión a la red. En un escenario de movilidad, un sistema final cambiará la dirección topológica del punto de conexión manteniendo sus direcciones IP y MAC inalteradas.

5.2.2 Comparación de coordenadas RSTAA con direcciones MAC e IP.

Las coordenadas RSTAA expresan conectividad jerárquica desde el origen de la red campus (Bridge Raíz). Tienen características comunes y diferenciadoras con las direcciones IP y con las direcciones MAC. En la Tabla 4 se ilustran las principales características en comparación con las direcciones IP y MAC.

Tabla 4. Comparación características direcciones MAC, IP y coordenadas RSTAA.

	Direcciones MAC	**Direcciones IP**	**Coordenadas RSTAA**
Longitud	48 / 64 bit	32 /128 bit	64 bit / variable
Unicidad	Sí	Sí	No Sí (1)
Validez	Mundial	Mundial	Local Mundial(1)
Jerárquicas/ Planas	Planas	Jerárquicas	Jerárquicas
Vinculación al HW	Por ítem (fábrica)	No	Sí (por número de puerto)
Asignación	IEEE/ Fabricante	Manual / Semiautom.	Automática
Identifica	NIC, item físico, fabricante	Conectividad lógica Subred	Conectividad física Posición topológica en el Árbol de Distribución
Adecuado para	NIC	Enlace Router	Enlace ARB
Utilizada por	Bridges/ Conmutadores/ Sistemas finales	Encaminadores	Adaptative Routing Bridges (opc.)
Estabilidad	Alta. Física	Media Administrativa (ISP)	Media Topológica
Actualización	Ninguna	Manual Semi autom. (DHCP)	Automática

(1): Unicas y universales si incluyen la dirección MAC del Bridge Raíz.

Las direcciones MAC son únicas, identifican la terminación (*endpoint identifier)* y las IP identifican jerarquía y posición *(locator)* de una interfaz en la red. Las coordenadas RSTAA son aún más del tipo *locator* que las direcciones IP por estar vinculadas a la

topología física, indican un camino de acceso a las mismas y se utilizan para identificar un nodo, no una interfaz. Su validez, en cambio, es de ámbito local, salvo si se combinan con la identidad del bridge raíz de la red, la cual contiene su dirección MAC, por lo que la combinación de Bridge ID y coordenada RSTAA será única y de ámbito universal.

Las coordenadas RSTAA contienen información topológica pero no *identidad*, al contrario que las direcciones MAC. La identidad se adquiere si se asocian con una dirección MAC o IP o bien con la identidad del Bridge Raíz. Por ello presentan ventajas operativas para el encaminamiento y direccionamiento dinámico.

5.3 Protocolo RSJ

En esta sección se describe el protocolo RSTAR Jerárquico (RSJ). El protocolo RSJ es una extensión jerárquica del protocolo RSTAR. Su modelo formal, resultado de extender y generalizar el modelo del protocolo STAR [Lui02], se describe en el Apéndice A, cuya referencia es necesaria para la comprensión de la descripción de RSJ en esta sección.

El protocolo RSJ se propone como protocolo opcional en el nivel inferior, de forma alternativa a RSTP y a RSTAR, cuando se desea una mayor escalabilidad de los tamaños de red inferior manteniendo la optimización de caminos en la red. Sus características de jerarquización mejoran la escalabilidad y su algoritmo de vectores distancia favorece la simplicidad y robustez. Las principales características del protocolo RSJ son las siguientes:

- Permite utilizar enlaces entre bridges RSJ que no forman parte del árbol de expansión, cuando su coste es inferior al del camino por el árbol de expansión, aumentando además la utilización de la infraestructura. Los enlaces redundantes entre bridges de un nivel y los del inmediato superior e inferior, que normalmente son desactivados por RSTP, pueden ser utilizados por RSJ aumentando el rendimiento y utilización de la red.
- Utiliza coordenadas RSTAA, basadas en el árbol de dispersión estándar RSTP, como auxiliar para el encaminamiento y el procesado de la topología de la red.
- Utiliza encaminamiento mediante vectores distancia jerárquicos y con radio de difusión limitado por lo que se adapta a redes relativamente grandes con sobrecarga reducida.
- Al igual que STAR, puede coexistir con bridges estándar 802.1D (RSTP) y extenderse de forma progresiva reemplazando bridges estándar por bridges RSJ.

En vez de una tabla de distancias plana utiliza una tabla jerarquizada, en la que las rutas a los nodos se agregan jerárquicamente con arreglo a las coordenadas RSTAA. En dicha tabla, cada entrada incluye:

- Distancias estimadas, máximas y mínimas a los bridges RSJ del área.
- Puerto de salida hacia bridges destino

- Identidad del bridge RSJ de siguiente salto hacia RSJ destino.
- Coordenada RSTAA del bridge destino a.b.d (en lugar de los indicadores definidos en STAR que identifican el rol del bridge: *Antecesor, Descendiente, Distinta Rama, Mismo Camino)*. (Ver Figura 51).

Al igual que STAR, RSJ difiere de los protocolos de Vector Distancia convencionales en que en ocasiones no es posible conocer la distancia exacta al bridge destino, en cuyo caso se utilizan las distancias estimadas máxima y mínima. El radio de difusión de rutas es limitado y dependiente del nivel jerárquico.

5.3.1 Procesos RSJ

Según el modelo de STAR y RSTAR descritos más arriba, RSJ incluye por una parte los procesos del protocolo RSTP más los específicos de RSJ como son la asignación de coordenadas RSTAA y el encaminamiento RSJ. Por tanto RSJ resulta de extender RSTAR con la asignación de coordenadas RSTAA y el encaminamiento RSJ.

Los procesos principales que componen RSJ se muestran en la Figura 49 y son:

- Elaboración del árbol RSTP, de acuerdo al estándar IEEE 802.1D.
- Asignación de coordenadas RSTAA, según se ha descrito en el protocolo RSTAA.
- Procesos estándar RSTP de reenvío de tramas y aprendizaje de direcciones MAC en puertos, de acuerdo al estándar IEEE 802.1D.
- Procesos básicos de aprendizaje y reenvío RSJ, descritos en los epígrafes siguientes.

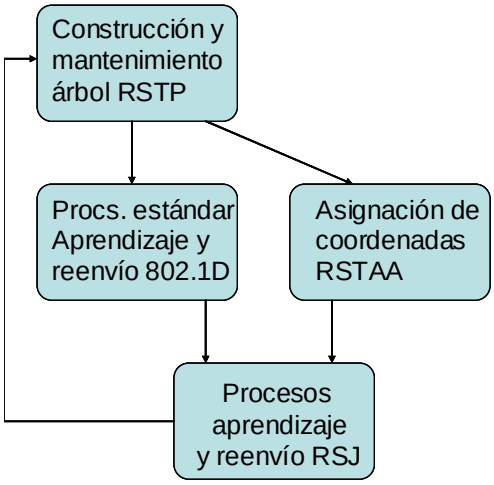


Figura 49. Procesos RSJ

5.3.2 Procesos de aprendizaje y reenvío RSJ.

Los procesos básicos de RSJ son:

- Proceso de Búsqueda de Caminos RSJ.

Este proceso calcula la tabla de encaminamiento del bridge RSJ (*TEB*). Esta tabla, si es posible, contendrá el bridge RSJ de siguiente salto y puerto de salida del camino más corto para cada bridge RSJ destino. A su vez consta de tres procesos:

  - Descubrimiento de Vecinos RSJ Directos. Su objetivo es descubrir a los bridges RSJ vecinos inmediatos y obtener y actualizar las distancias a ellos.
  - Descubrimiento de Vecinos RSJ Distantes. El objetivo de este proceso es descubrir otros bridges RSJ que están conectados a través de bridges no RSJ y obtener distancias exactas o aproximadas a ellos.
  - Mejora de Vector Distancia. El objetivo es obtener rutas jerárquicas a todos los bridges RSJ. Se describe en epígrafe aparte.

- Proceso de Aprendizaje RSJ.

En el protocolo STAR original, este proceso consiste en el aprendizaje de los sistemas finales asociados a cada bridge (de los que el bridge es *designado*) mediante intercambio de listas de sistemas finales. En RSJ, por utilizarse servidores ARP, no es necesario intercambiar estas listas de sistemas finales, dado que el servidor ARP facilita el AB designado del host destino.

- Proceso de Reenvío RSJ.

Este proceso consiste en el envío de tramas entre dos bridges RSJ utilizando caminos alternativos encontrados mediante el protocolo RSJ. Utiliza la tabla de reenvío (forwarding) calculada por el proceso *Búsqueda de caminos.*

- Resolución ARP, AB y registro.

El proceso de resolución ARP, resolución de AB y registro del sistema final para ARP se ha descrito en detalle en el epígrafe 4.3. La diferencia en el protocolo RSJ respecto al protocolo RSTAR es que el servidor ARP/AB al que el AB reenvía el paquete ARP del sistema final, incluye en la contestación la coordenada RSTAA del bridge AB agente del host destino, que es utilizada en el encaminamiento RSJ en la red inferior. Como se utilizan servidores ARP/ABs para obtener la dirección MAC del sistema final destino y del AB destino, el aprendizaje de sistemas finales por parte de cada AB se limita a los SF de los que el AB es bridge agente.

5.3.3 Operación del protocolo RSJ

5.3.3.1 Establecimiento de coordenadas RSTAA a vecinos y descubrimiento de vecinos directos RSJ.

La asignación de coordenadas RSTAA forma parte del descubrimiento de bridges vecinos directos RSJ y se produce al terminar la construcción del árbol de expansión usando RSTP. Periódicamente los bridges RSJ envían, por los puertos designados, además de las BPDUs RSTP, BPDUs con el valor “protocolo RSJ” en el campo tipo de la trama Ethernet (en vez del valor 0x42h correspondiente al protocolo de árbol de

expansión), tramas que contienen la coordenada RSTAA que el bridge RSTAA antecesor asigna al bridge RSJ que esté conectado a ese puerto designado. Solamente los bridges RSJ vecinos inmediatos las procesan y obtienen así su coordenada RSTAA. Los bridges 802.1D que ejecutan STP o RSTP descartan este tipo de tramas.

Mediante las tramas estándar RSTP el bridge RSJ conoce a los bridges vecinos y el coste de sus enlaces. Sabe cuáles de ellos son RSJ porque recibe las BPDUs RSJ conteniendo la coordenada RSTAA, coste al bridge raíz y coste a su bridge padre (BPDU *MiInfo).*

5.3.3.2 Descubrimiento de vecinos distantes RSJ.

Cada bridge RSJ precisa descubrir a los bridges RSJ vecinos porque muchos no son vecinos directos por existir algún bridge 802.1D interpuesto en el camino.
El bridge RSJ trata de encontrar a los bridges RSJ del árbol de expansión que están separados por uno o más bridges 802.1D y estimar la distancia a ellos. Los bridges estándar reenvían y no procesan estas tramas de descubrimiento y los bridges RSJ procesan y no reenvían estas tramas de descubrimiento.

Este proceso funciona de la forma siguiente: los bridges RSJ envían, a través de su puerto raíz, si el bridge conectado al puerto raíz no es RSJ, BPDUs de tipo *MiInfo* a la dirección Multicast *"todos los Bridges RSJ"* indicando su coordenada RSTAA (el proceso de asignación de coordenadas se detalla en un epígrafe posterior). Esta BPDU es reenviada por todas las ramas del árbol de expansión a partir del puerto raíz hasta encontrar un bridge RSJ en cada segmento. Este bridge recibirá a su vez las BPDU *MiInfo* de los situados debajo de él en el árbol de expansión de la misma forma. Un ejemplo se muestra en la Figura 50.

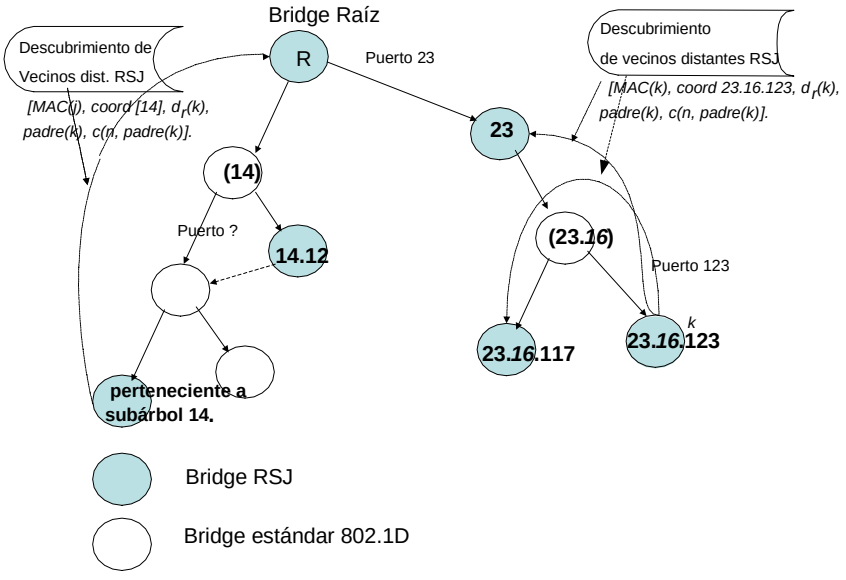


Figura 50. Descubrimiento de vecinos distantes RSJ

Por lo tanto, cada bridge RSJ encontrará solamente un bridge RSJ *antecesor* como máximo y ninguno, uno o varios bridges RSJ *descendientes* (ver definiciones en Apéndice A). Las BPDUs usadas son de dos tipos: *MiInfo*, en la que el bridge RSJ envia

su coordenada RSTAA, la distancia a su bridge antecesor RSJ y la distancia al bridge raíz; y *NuestraInfo,* con la que un bridge RSJ comunica a otro bridge RSJ la distancia vía árbol entre ambos cuando la conoce. Como no siempre se conoce esta distancia, esta segunda BPDU no se utiliza siempre. Mediante las coordenadas RSTAA, cada bridge puede hacer alternativamente una estimación bastante buena basada en las profundidades de coordenadas de ambos bridges y en la distancia propia al bridge raíz (conocida).

El formato de *MiInfo* es: [*MAC(k), coord(k),* $d_r(k)$*, padre(k), c(n, padre(k)*]. *MAC(k)* es la dirección MAC del bridge, *coord(k)* es la coordenada RSJ del bridge, $d_r(k)$ es su distancia via árbol al bridge raíz, *padre(k)* es la dirección MAC de su bridge padre, sea o no RSJ y *c(n, padre(k))* es la distancia a su bridge padre. La trama va encapsulada con la dirección de destino de grupo multicast "todos los bridges RSJ". La BPDU *NuestraInfo* contiene las identidades de los dos bridges y la distancia vía árbol entre ellos. La BPDU *MiInfo* se procesa de la forma siguiente:

- Se comparan las coordenadas RSTAA de origen y destino.
- Se deduce la distancia vía árbol. La distancia vía árbol es la del camino a través del bridge común más bajo que aparece en sus coordenadas RSTAA. Si las coordenadas RSTAA de los bridges son *a.m.n* y *a.l* el bridge común situado más abajo en el árbol (*antecesor común más cercano)*, es el bridge *a.* Si los bridges están situados en ramas distintas del árbol de expansión, las coordenadas no tienen ningún elemento común y el único bridge común es el bridge raíz.
  - Como resultado de la comparación de las coordenadas pueden aparecer tres casos:
    - El bridge está en otra rama y no es *hermano* (no tiene el mismo padre común).
      - En este caso el proceso consiste en estimar una horquilla de distancias mediante una máxima y mínima:
        - Máxima. Se utiliza como distancia máxima la suma de las distancias respectivas de ambos bridges al bridge raíz. Esta distancia se reducirá cuando se encuentra a un bridge antecesor común a ambos, reemplazando la nueva distancia vía árbol suma de las distancias hasta el antecesor común a la distancia anterior vía bridge raíz.
        - Mínima. La distancia mínima se estima como la suma de las distancias a los bridges "padres" (designados) de ambos. El flag de *Árbol* se carga con el valor *camino_de_árbol* y el flag de *Antecesor* a *Nulo.*
    - El bridge es *hermano*.
      - Este caso se detecta porque ambos bridges tienen el mismo padre.
      - En este caso el bridge hermano está en el árbol de expansión.
      - El proceso consiste en fijar la distancia como exacta porque la distancia de árbol coincide con la real.
      - Se anota en tabla el puerto por donde se recibio la trama.
    - El bridge es *antecesor*.
      - El proceso en este caso consiste en fijar la distancia como exacta (la distancia de árbol coincide con la real).
      - Se anota en tabla el puerto por donde se recibio la trama.

5.3.3.3 Procesado de Vecinos Directos

Cada bridge RSJ conoce la distancia a sus bridges RSJ vecinos directos a los que está conectado por la información topológica obtenida al establecer y mantener el árbol de expansión. Por ello este, proceso no precisa utilizar BPDUs adicionales. Los vecinos directos RSJ pueden ser vecinos de árbol o no. En la Figura 77 del Apéndice se muestran ejemplos de bridges *vecinos de árbol* y *no vecinos de árbol*. Los enlaces entre bridges no vecinos de árbol son *enlaces cruzados utilizables (elegibles)* mientras que los enlaces cruzados entre vecinos de árbol, entre nodos que pertenecen a una misma rama del árbol de expansión, no son utilizables o elegibles.

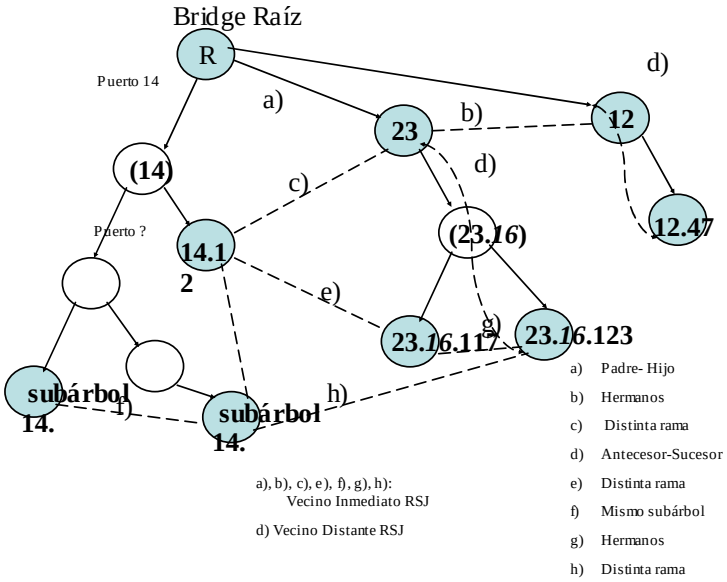


Figura 51. Relaciones topológicas entre bridges RSJ.

Como resultado de la comparación de las coordenadas son posibles cuatro casos:

- El bridge vecino es vecino de árbol (pertenece a la misma rama del árbol de expansión, no es enlace cruzado elegible). Los casos posibles son:
  - Es hijo. La distancia es exacta, se rellenan las distancias en tabla *dmin = dmax = c(n,k),* el camino es a través del AE.
    Flag A = Árbol, Flag R =Descendiente
    - *siguiente_salto=K (hijo), puerto forwarding= puerto designado para el hijo k*
  - Es padre. La distancia es exacta, *dmin=dmax= c(n,k), e*s a través del AE.
    Flag A =Árbol, Flag R =Antecesor
    - *siguiente_salto=k (padre), puerto forwarding= puerto raíz del bridge*

- El vecino no es vecino de árbol (es enlace cruzado *utilizable o elegible):*
  - Es además bridge vecino RSJ distante.
    - En este caso existe ya una entrada en la tabla para este bridge obtenida en el proceso de vecinos distantes.
    - En este caso si la distancia existente en la tabla (DT) es exacta se elige la menor distancia entre la del enlace cruzado la de árbol.
    - Si DT es aproximada, solamente se sustituye si la distancia es menor que la estimación mínima (límite inferior de la distancia)
  - No es bridge RSJ distante. En este caso no figura en la tabla DT por lo que es insertado con su distancia.

5.3.3.4 Mejora de Vectores Distancia.

Una vez que se completa el procesado de vecinos directos, los bridges RSJ intercambian sus VDs en el procedimiento llamado Mejora de Vectores Distancia. Este proceso es semejante al de un protocolo de vectores distancia. De esta forma descubren bridges RSJ no vecinos y rutas hacia ellos. Cuando la ruta descubierta es más corta que la conocida, la reemplaza.

Las BPDUs utilizadas se llaman VD_{RSJ}, se transmiten encapsuladas con la dirección MAC multicast de grupo de los bridges RSJ y contienen lo siguiente:

[*bridge origen , bridge destino, RSJorigen, RSJdestino, distancia mínima, distancia máxima, flag A árbol, flag R (antecesor/sucesor), TTL***]**

RSJdestino y *RSJorigen* se utilizan como prefijos de encaminamiento. El protocolo RSJ no calcula ni transmite rutas a todos los bridges RSJ sino solamente a los de su radio de acción, que se derivan de su nivel jerárquico según coordenadas del bridge. Las tramas recibidas se encaminan al bridge RSJ que tenga la coincidencia de prefijo más larga de coordenada con el bridge destino. De no existir, o ser más largo que la distancia de árbol, se encaminan via árbol de expansión. El proceso es el normal de un protocolo de vector distancia con algunas especificidades:

- Al recibir un nuevo vector distancia, se recalculan las rutas y las que se vean reducidas se modifican de la forma correspondiente cambiando el siguiente salto al nodo que envió dicho vector distancia. Se añade la ruta con coordenadas RSTAA
  - Si el nodo es nuevo en la tabla de encaminamiento
    - Se calculan la distancias: $d_{min}(k, n´) = d_{min}(k, n) + d_l$; $d_{max}(k, n´) = d_{max}(k,n) + d_l$
    - *siguiente_salto(k, n´) = n*
    - *Puerto_salida(k, n´)= Puerto_salida(k,n)*
    - *Flag Camino_de_árbol=Árbol si ambos flags Camino_de_árbol= Árbol*
    - *Flag R = Antecesor si ambos flags (COORD) = Antecesor, Descendiente si ambos =Descendiente, Nulo* en los demás casos.
  - Si el nodo no es nuevo (existe la ruta a ese nodo en la tabla de encaminamiento) En este caso hay dos posibilidades: la nueva ruta descubierta es camino de árbol o no lo es. Lo es si tanto el camino al vecino que comunica la ruta, como la ruta comunicada lo son (el flag de árbol está activado). Se comprueba con los flags camino_de_árbol (ambos en valor árbol) o por comparación de coordenadas.
    - La nueva ruta es un camino del árbol de expansión. Por ser la ruta nueva camino del árbol de expansión, solamente hay que verificar que el camino es más corto o igual que la ruta existente. Se asume la nueva ruta por el árbol si la distancia mínima (limite inferior) de k a n´ mas el coste del enlace es menor o igual que el limite superior de la ruta existente. Puede ocurrir que el camino por el árbol, por ser estimado, sea peor que la ruta anterior, pero se elige asi para evitar que el algoritmo elija una ruta peor que la existente en el árbol de expansión.
      - La vieja ruta no es camino del árbol de expansión. El criterio general que se aplica es reemplazar con la nueva ruta cuando es estadísticamente mejor según se indica más abajo.
        - Para mantener la condición de que no se empeora el camino, el límite superior de la nueva ruta no debe ser superior al de la ruta existente.

        - Para la condición de estadísticamente mejor, se compara el coste promedio ((max+min)/2) de ambas rutas y debe cumplirse que el coste promedio de la nueva sea menor o igual que el de la existente. Si el promedio es igual, la diferencia entre límites debe ser menor que la de la existente.
- La nueva ruta NO es un camino del árbol de expansión. Los casos son:
    - La vieja ruta es camino del árbol de expansión.
        - El criterio es reemplazar solamente cuando la nueva ruta es *absolutamente mejor* que la existente vía árbol.
            - Para ello se compara el límite inferior de la ruta existente con el límite superior de la ruta nueva.
            - La vieja ruta no es camino del árbol de expansión. Se elige el camino estadísticamente mejor, asegurando que nunca empeora el del árbol, aun a costa de perder mejoras potenciales.

Como puede verse, el procesado depende de la posición relativa entre bridges RSJ, la cual se obtiene de forma directa por comparación de las coordenadas RSTAA. De la comparación de coordenadas entre el bridge origen *o* y destino *d* se obtiene lo siguiente:

Definimos *nivel(B)=n* donde B tiene la coordenada a.b.c donde *n* es el número de enlaces por el AE desde el bridge B hasta el bridge raíz.

Defininimos *coord (B)=a.b.c* coordenadas del bridge B (10 bit por coordenada, 0 a 1023, correspondiente al número de puerto del bridge)

- Definimos *coord(d) AND coord(o) = acc(o,d)* (antecesor común más cercano**)**

Si dos bridges tienen coordenadas a.b.c.d y a.b.e.k.l, el antecesor común más cercano es :

- *acc(a.b.c.d, a.b.e.k.l) = a.b*

La relación de antecesor-sucesor se determina de la forma:

- *Si coord(d) AND coord(o) = coord(o), o* es *antecesor* de *d*
- *Si coord(d) AND coord(o) = coord(d), o* es *sucesor* de *d*

5.3.4 Difusión de coordenadas y distancias

5.3.4.1 Difusión de rutas limitada

La difusión de vectores distancia puede limitarse. Dado el carácter complementario del encaminamiento RSJ respecto al Árbol de Expansión, no es necesario difundir todos los vectores distancia a todos los nodos, evitando así la sobrecarga típica de los protocolos de vector distancia. La jerarquía de coordenadas permite sin embargo mantener rutas alternativas a la mayoría de nodos. El radio *r* de difusión de cada ruta es función del nivel *n* del bridge que la anuncia: $r = m*(n_{max}-n)$ (siendo *m* configurable). Con *m* se modula la profundidad de difusión de las rutas (mayor interconectividad transversal y oblicua respecto a Árbol de Expansión). Para *m* igual a 2, el radio de difusión disminuye en 2 en cada nivel jerárquico descendido. Para $n_{max} = 5$ los radios resultantes son: para nivel 1, 8; para nivel 2, 6; para nivel 3, 4 y para nivel 4, 2. La limitación de difusión no requiere necesariamente configuración, puede preconfigurarse

en el propio protocolo a valores optimizados para determinados tamaños de redes campus.

La profundidad máxima n_{max} del árbol de expansión es difundida por toda la red mediante las BPDUs de direccionamiento RSJ enviadas por cada bridge RSJ por su puerto raíz. Cada RSJ envia hacia el bridge raíz su coordenada RSJ asignada. El bridge raíz difunde en broadcast periódicamente la profundidad máxima n_{max}.

5.3.4.2 Cancelación de coordenadas y convergencia

Se hacen a continuación unas consideraciones sobre la estabilidad y convergencia del protocolo RSJ. El empleo de RSTP como encaminamiento por defecto a través del árbol de expansión hace inherentemente estable a RSJ. Cuando se producen cambios de topología en la red se utiliza solamente el encaminamiento a través del árbol de expansión para asegurar la convergencia rápida. Al detectarse un cambio de topología como la caída de un enlace, RSTP borra las direcciones MAC aprendidas que pueden verse afectadas por dicho cambio de topología al recibirse dicha notificación en un determinado puerto. En cuanto al encaminamiento complementario de RSJ, su estabilidad está ligada a la de las coordenadas RSTAA utilizadas. Las coordenadas de un bridge dejan de ser válidas cuando algún bridge antecesor cambia su coordenada. Los cambios de topología producidos por la caída de un bridge o de un enlace pueden producir reconfiguraciones. El protocolo RSTP realiza transiciones rápidas acordadas localmente entre bridges vecinos para habilitar enlaces deshabilitados existiendo enlaces preasignados, como es el caso del rol de puerto denominado *alternate port* (conectado a través de un puerto designado de distinto bridge al mismo bridge raíz, puerto alternativo para ser elegido puerto raíz), mediante el que el protocolo puede elegir rápidamente el nuevo puerto raíz del bridge. Mediante éste y otros mecanismos los tiempos de reconfiguración son mayoritariamente del orden de pocas centenas de milisegundos. El mayor cambio de coordenadas RSTAA se produce cuando el bridge raíz queda fuera de servicio y debe elegirse otro bridge raíz.

Los cambios de topología de RSTP se difunden por toda la red de forma inmediata por todo el árbol de expansión con TCN en el octeto de indicadores de las BPDUs. Estas BDPU indicando cambio de topología producen, al difundirse por la red, el borrado inmediato de las tablas de encaminamiento RSJ existentes (además de las MAC aprendidas), retornando en todos los bridges al encaminamiento estándar básico a través del árbol de expansión. Una vez recuperado el encaminamiento RSTP, el protocolo reinicia el cálculo de las tablas de encaminamiento y se restablece el encaminamiento RSJ suplementario por enlaces cruzados entre bridges RSJ. El mecanismo de difusión de las BPDUs de cambio de topología de RSTP asegura la difusión más rápida posible.

5.3.4.3 Proceso de Reenvío

El proceso de reenvío de tramas en los bridges RSJ se muestra en la Figura 52. Es de la forma siguiente: El encaminamiento por defecto de las tramas es el del árbol de expansión RSTP establecido en la red. En cada bridge RSJ se comprueba si existe mejor ruta que la de árbol vía un enlace cruzado a otro bridge RSJ. El seudocódigo es el siguiente:

- El AB agente recibe del servidor ARP el AB agente del SF destino (AB destino) y su coordenada RSJ.
- Se analiza el AB destino:

If $coordAB_{dest}$ *AND* $coordAB_{este\_bridge} \neq nulo$ *;* el AB destino está en la misma rama.

If $coordAB_{dest}$ *AND* $coordAB_{este\_bridge} = coordAB_{des}$; AB destino está por encima del bridge actual,

Encapsular = False

puerto forwarding= puerto raíz(este bridge)

If $coordAB_{dest}$ *AND* $coordAB_{este\_bridge} = coordAB_{este\_bridge}$ *;* el AB destino está por debajo del bridge actual,

Encapsular = False

puerto forwarding= puertos designados(este bridge) ; o puerto aprendido.

Else ; el AB destino está en otra rama

If Tabla Encam(ABdestino) $\neq$ *nulo* ; hay ruta RSJ

Encapsular = True

ABsigte= TE(ABdestino).sigteAB, puerto= TD(ABdestino).puerto_sal

Else ; no hay ruta RSJ , encaminamiento por el árbol de expansión hacia el raíz.

Encapsular = False

puerto forwarding= puerto raíz(este bridge)

FI

FI

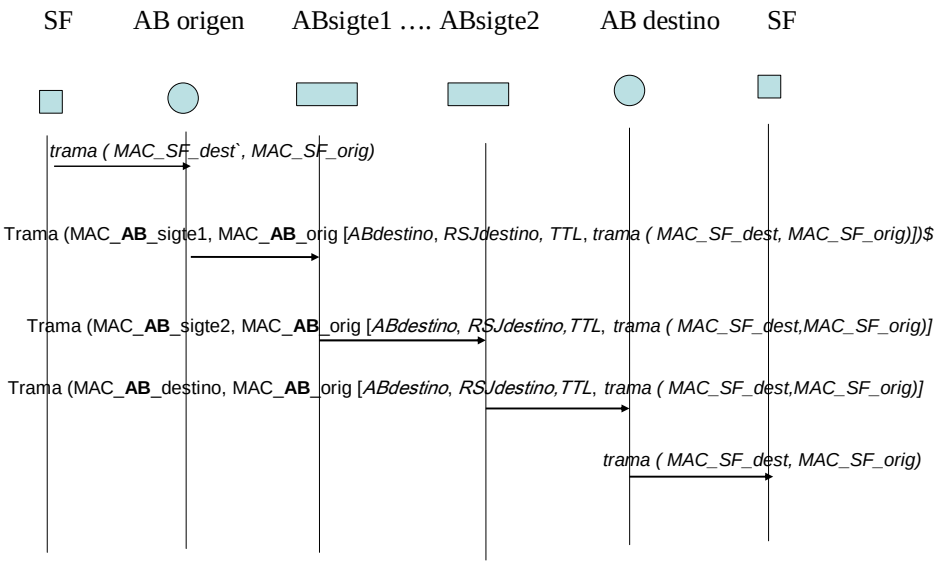


Figura 52. Escenario de encaminamiento RSJ.

En cada bridge RSJ atravesado el procesado es el mismo, reemplazando la dirección del bridge AB de siguiente salto en cada salto, mientras que en los bridges 802.1D la trama se difunde de la forma estándar, hasta alcanzar el AB destino. En el AB destino se desencapsula la trama y se envía al SF destino aprendido por el AB destino.

5.3.4.4 Duplicación de tramas

Un problema analizado en [Lui02] es el posible peligro de duplicación de tramas, si una trama fuera enviada por el árbol de expansión y por un enlace cruzado entre bridges STAR. Por las mismas razones que se exponen allí no existe este peligro de duplicación de tramas dado que el AB Agente es el único bridge que efectúa el reenvío de tramas y lo realiza por el enlace cruzado cuando procede o por el árbol de expansión cuando el encaminamiento alternativo no es favorable, pero nunca simultáneamente.

5.3.4.5 Descarte de tramas

Al igual que en el protocolo STAR y por los mismos motivos, no existe el peligro de descarte de tramas descrito (una trama destinada a un bridge situado en otra rama es enviada en sentido descendente (en vez de ascendente) por esa rama del AE, al atravesar un bridge intermedio es descartada por recibirse por el puerto raíz de ese bridge (puerto por el que se habrá aprendido esa dirección). El procedimiento utilizado es encapsular la trama de datos entre el bridge RSJ del árbol y el siguiente bridge RSJ que actúa como *proxy*, según se describe en [Lui02].

5.3.5 Protocolo RSJ en Redes Mixtas.

El protocolo RSJ tiene rendimiento óptimo cuando todos los bridges ejecutan dicho protocolo. En esas condiciones, todos los bridges tienen una coordenada RSTAA distinta. Si no todos los bridges son RSJ puede no ser posible asignar coordenadas RSTAA distintas a todos los bridges, compartiendo distintos bridges de un subárbol una misma coordenada. Es importante sin embargo que RSJ funcione en redes mixtas con bridges estándar que solamente ejecutan RSTP. En esta sección se describe el funcionamiento propuesto del protocolo RSJ en redes mixtas. En este caso pueden algunas áreas (grupos de bridges conectados como una rama del Árbol de Expansión) obtener una coordenada conjunta RSTAA utilizable para el encaminamiento si la cadena de asignación de coordenadas RSTAA es discontinua en dos o más bridges 802.1D consecutivos.

5.3.5.1 Coordenadas RSTAA en redes mixtas

La asignación automática de coordenadas basadas en árbol de expansión que realiza RSTAA no funciona cuando uno de los bridges del árbol de Expansión no ejecuta el direccionamiento RSTAA, creando una discontinuidad en la cadena de asignación de coordenadas desde el bridge raíz en sentido descendente. Es posible, sin embargo, bajo los siguientes supuestos, realizar la asignación de coordenadas RSTAA:

-Puede haber bridges 802.1D RSTP entre bridges RSTAA, pero solo serán operativas las coordenadas jerarquicas para encaminamiento RSJ si el número de bridges entre RSTAAs es $n<=2$. Puede contemplarse el que el RSTP ejecutado por los RSJ en igualdad o diferencia de coste menor que X (porcentaje del coste) elija como puertos raíz los que estan conectados a bridges RSJ (ver más abajo), al precio de alterar ligeramente el comportamiento estándar de formación de árboles de expansión en los bridges RSJ.

-Si n=1, las coordenadas se pueden asignar normalmente, dado que el bridge RSJ que recibe BPDU RSJ de un bridge situado por encima de él, puede generar su coordenada

RSTAA añadiendo a la coordenada RSTAA transmitidad por el bridge RSJ superior la ultima coordenada obtenida del protocolo RSTP (de las BPDUs de su bridge RSTP padre) en lugar de la forma inicialmente prevista a traves del protocolo RSTAA específicamente. Ejemplo 23.12 de la Figura 53.

-Si n=2, se asignará una coordenada RSTAA del tipo 23.X.11 y el RSTAA que asigna la dirección no asignará ninguna coordenada más en esa rama, para evitar tener que gestionar la bifurcación de coordenadas. La coordenada X indica la existencia de dos bridges 802.1D contiguos en la rama del árbol de expansión y por tanto el bloqueo de la asignación de coordenadas.

Pueden pues coexistir coordenadas 23.15 con una coordenada 23.X.15, y con ilimitadas coordenadas del tipo 23.16.117, 23.16.12, etc. (El bridge 23.16 es único aunque no ejecute RSTAA).

De esta forma, si dos bridges RSTP (no RSJ) están conectados consecutivamente en el árbol de expansión, ese subárbol completo bajo los dos bridges queda inhabilitado para mejoras de caminos RSJ y todos los bridges RSJ de el subárbol se direccionan como "subárbol a.b.d" donde a.b.d corresponde a la dirección del primero de los dos bridge s no RSJ. En la Figura 53 se muestra cómo los bridges inferiores quedan con coordenadas subárbol 14 mientras que el bridge 14.12 obtiene coordenada exacta RSTAA por estar separado solamente por un bridge estándar de su bridge RSJ antecesor R en el árbol de expansión. El bridge raíz debe ser necesariamente RSJ para poder asignar las coordenadas. Todos los bridges deben ejecutar RSTP, algunos de ellos además RSJ, que desde el punto de vista de implementación incluye el protocolo RSTP.

El protocolo puede configurarse de forma que el establecimiento del árbol de expansión prevenga estas situaciones en la elección de puerto raíz (normalmente se elige el puerto por el que se recibe menor coste al bridge raíz), de forma que los bridges RSTAA unidos a bridges 802.1D no los elijan salvo desventajas significativas de coste de camino (si la diferencia de coste es menor de una cantidad o un porcentaje relativo del coste). El algoritmo de elección de puerto raíz debe modificarse en los bridges RSJ de la forma:

- *If este_bridge= RSJ and $c_r(p_{cr})$ < min(..$c_r(p_i)$...) and bridge_conectado_a(p_{cr})=RSJ ; ambos bridges son RSJ*
  - *select p_{cr}= puerto_raíz*
- *else if (este_bridge=RSJ or (bridge_conectado_a(p_{cr})=RSJ)) and ($c_r(p_{cr})$ < k* min($c_r(p_i)$)) ; uno de ellos al menos es RSJ*
  - *select p_{cr}= puerto_raíz*
- *else if ; caso de dos bridges no RSJ contiguos:*
  - *$c_r(p_{cr})$ < k´* min $c_r(p_i)$; Si hay algun puerto conectado a bridge RSJ cuyo coste sea menor que k´ *min $c_r(p_i)$*
  - *select p_{cr}= puerto_raíz*
- *else*
- *eleccion normal puerto raíz según 802.1D*
- *end*

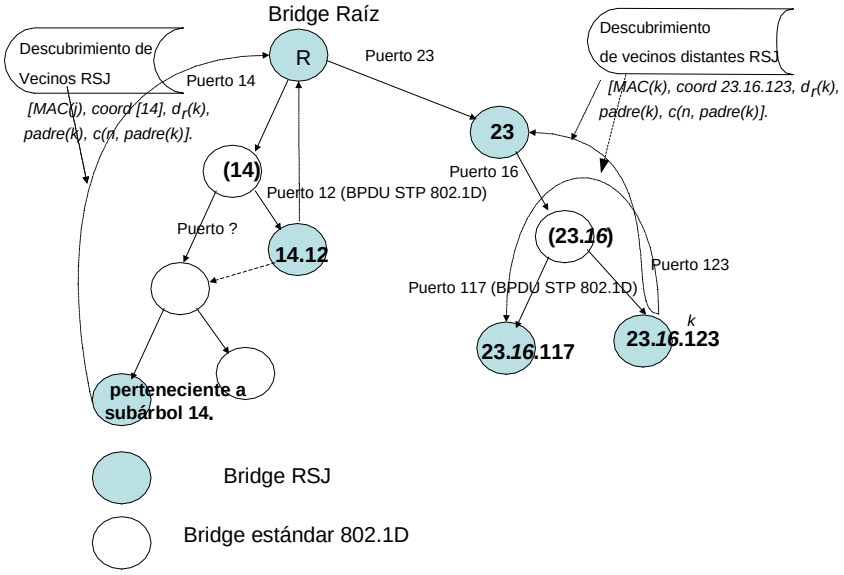


Figura 53. Asignación de coordenadas RSTAA en redes mixtas.

Los factores de ponderación *k* y *k´* permiten distintas ponderaciones de forma que la elección del segundo bridge no RSJ se realice cuando no hay otra opción.

5.3.6 Propuesta ARP mediante servidores en reparto de carga.

En la arquitectura propuesta, para resolver las direcciones de capa 2 es necesario obtener, a partir de la dirección IP del sistema final destino, la direcciones MAC del AB destino. En el caso de tráfico ínter-área, el ARB del área (origen) resuelve el ARB destino a partir del AB destino. Existen múltiples soluciones alternativas para limitar la difusión del protocolo ARP en las redes campus que proponemos. De entre ellas el mecanismo de ARP basado en servidores propuesto ha sido descrito con detalle en la sección denominada Escenario de Encaminamiento por lo que no lo repetiremos aquí en los aspectos ARP, sino que describimos a continuación el protocolo para el reparto de carga entre servidores.

El mecanismo basado en servidores ARP/AB en reparto de carga por direcciones IP está orientado a repartir de forma unívoca la carga de resolución ARB según direcciones IP. Los servidores ARP se comunican entre sí por la dirección ARP multicast pero dan servicio por su dirección IP unicast exclusiva. El AB recibe la trama *ARP (IPdest, MACdest?, IPorig, MACorig*). Calcula *hash (IPdest)* de pocos bits (2,3,4) según el numero de servidores ARP activos previsto para atender el servicio (tantos como resultados hash distintos: 4, 8, 16 respectivamente). El cálculo del hash puede realizarse mediante un simple circuito de generación de CRC. Los servidores ARP activos anuncian en la dirección de grupo multicast de los *servidores ARP* activos su dirección MAC, IP y las *hash ID* bajo su responsabilidad. Cada AB recibe, mediante suscripción al grupo

multicast de los se*rvidores ARP activos* el/los valor/es de *hash (IP)* que atiende cada servidor/registrador y sus direcciones MAC e IP unicast para ser consultado y para registrar sistemas finales. Opcionalmente se pueden habilitar mecanismos dinámicos entre los servidores para repartir la carga entre los servidores activos, cediendo parte de los grupos de direcciones hash (IP) a los servidores a medida que se activan. Un ejemplo es el siguiente:

El primer servidor/registrador ARP/AB que arranca posee las 16 (8, 4, según el número de bits empleado en el hash) combinaciones *hash IDs* y cuando van arrancando otros servidores cede una *hash ID* a otro en cada intervalo, cesando el registro de nuevos sistemas finales correspondientes a esa *hash ID*, dejando caducar los que tenia registrados, momento en el que deja de responder a las peticiones de resolución de dicha *hash ID*.

El servidor/registrador ARP/AB que recibe una *hash ID* cedida de otro, reenvía las preguntas (*who am I)* de los sistemas finales que no es capaz de resolver por no tener entrada en registro, al anterior servidor ARP encargado de esa *hash ID* (o alternativamente hace multicast a los servidores ARP para resolverlo) hasta que expira el intervalo de caducidad de dicho servidor.

Cuando un servidor/registrador ARP/AB cae y no puede dar servicio por cualquier razón, el AB espera la finalización de su tiempo de espera de respuesta y realiza un ARP estándar para efectuar la resolución. El funcionamiento convencional de ARP mediante difusión debe mantenerse en los ABs como mecanismo alternativo de emergencia para que la red sobreviva a una caída general de los servidores ARP. No se ha estudiado la posible aplicación de algoritmos de hashing robusto, por no ser aspecto prioritario en la Tesis y el que dada la caducidad de los datos no parece a primera vista efectivo complicar el algoritmo.

Entre los servidores ARP, que conocen las *hash ID* atendidas por cada servidor pueden habilitar un mecanismo unívoco de elección del servidor que tomará a su cargo la resoluciones de la *hash ID* del servidor que ha fallado. El más simple, pero seguramente no el más idóneo consiste en elegir el servidor que tenga dirección MAC más baja.

5.3.7 Otras alternativas para ARP

Como alternativas al reparto de carga descrito, pueden aplicarse otras soluciones más o menos convencionales. En la solución convencional, los paquetes de interrogación ARP son enviados en broadcast a todos los sistemas finales de la red, siendo procesados por ellos y contestados solamente por el sistema final interrogado. Esta opción supone una sobrecarga probablemente inaceptable para los sistemas finales en el caso de redes de gran tamaño, pero son posibles ciertas optimizaciones, como la consistente en que los ABs o ARB realicen la función de Proxy ARP, en todas o algunas ocasiones. Estas optimizaciones no están exentas de inconvenientes [Per05].

5.3.8 Neighbor Discovery (IPv6)

A diferencia del protocolo IP versión 4, en el que se utiliza ARP para resolver direcciones IP a MAC, IP versión 6 no utiliza ARP sino el protocolo de descubrimiento de vecinos (Neighbour Discovery Protocol (ND)). El protocolo de Neighbour Discovery agrupa varias funcionalidades: de descubrimiento entre sistema final y encaminador, de comunicación entre sistemas finales y la función de redirección (*redirect).* Las funcionalidades de descubrimiento son las siguientes: descubrimiento de encaminador, de prefijo, de parámetros y auto configuración de dirección. Las funciones de comunicación entre sistemas finales son: Resolución de direcciones, determinación de siguiente salto, detección de vecino inalcanzable y detección de direcciones duplicadas.

El protocolo utiliza paquetes ICMP que se envían a la dirección multicast en lugar de realizarse broadcast, lo que evita el impacto producido por la difusión de ARP en IPv4.

Como ejemplo, para resolver una dirección de capa 2, el sistema final envía un mensaje de Neighbor Solicitation (ICMP 135) a la dirección FF02:1 link-local con su dirección link local. La contestación se envía como Neighbor Advertisement (ICMP 136). Dado que no puede suponerse incluída siempre la dirección de capa 2 en la dirección IPv6, es necesario utilizar siempre Neighbor Discovery y no puede obtenerse de la dirección IP destino. El procesado propuesto para los paquetes de Neighbor Solicitation es en principio el mismo que para ARP. El hipotético despliegue masivo de IP versión 6 podría plantear tratamientos alternativos.

5.4 Tablas y estructuras de datos

Se describen a continuación las principales estructuras de datos empleadas por los protocolos propuestos: AMSTP y RSJ. Se describen en los apartados sucesivos agrupadas por tipo de dispositivo. Las estructuras de datos originales de STAR, de las que derivan las de RSJ están descritas en [Lui02].

5.4.1 En ARBs

Tabla Localización de ABs (TLA)
Esta tabla contiene la lista que asocia los ABs con su ARB padre o cabecera del árbol. Es una lista de tuplas con contenido: (identidad de AB (MAC), identidad de ARB (MAC)).

Base de Datos de Reenvío (BDR) *Forwarding Database (FD).* Esta tabla contiene las direcciones MAC aprendidas por el ARB como bridge. Incluye normalmente las direcciones MAC de los dispositivos activos conectados al ARB de forma directa o remota dentro del núcleo, de la red inferior y las configuradas manualmente (estáticas).
Tuplas: (ARB(MAC), host (MAC), puerto del bridge, tipo de puerto (puerto de troncal o de distribución). Por defecto, las direcciones MAC en las tablas caducan cada 300 segundos si no se refrescan.

Árbol RSTP
Es el elemento base en el procesado del protocolo RSTP. El estado del protocolo en cada puerto se guarda como *Vector de prioridad de puerto.* Contiene, por cada puerto:

- *ID Bridge raíz.*
- *Coste interno al raíz.*
- *ID Bridge designado.*
- *ID de puerto designado.*
- *ID puerto de receptor.*

El consumo de memoria es independiente del número de nodos, como es propio de los árboles de expansión.

Árboles Múltiples (AMSTP)
Por cada instancia de árbol (una por cada bridge ARB frontera) se almacena:

-Identidad de instancia (ligada a MAC ARB origen), MAC ARB origen, puertos pertenecientes a la instancia de árbol, roles (raíz o designado). En cada puerto se crea un vector de prioridad de puerto para cada instancia de árbol.

- *ID Bridge raíz de la instancia*
- *Coste interno al raíz de la instancia de árbol*
- *ID Bridge designado*
- *ID de puerto designado*
- *ID puerto receptor*

Las instancias de árbol resultan en una tabla de encaminamiento de la forma:
numero de instancia, MAC del bridge origen del árbol (raíz), ID puerto raíz, ID puerto designado 1....puerto designado n.

5.4.2 En ABs (bridges RSJ)

Los bridges AB utilizan las siguientes estructuras de datos:
TSF. Tabla de sistemas finales. Cada tupla contiene: *ID puerto, MAC(i),… MAC(k).* Siendo *MAC(i)* la dirección MAC del host *i* del que AB es bridge agente.

Por bridge RSJ destino
De acuerdo con la nomenclatura de STAR, denominamos *TVD(o,d)* a la tupla de información almacenada en el *o* relativa al camino de *o* a *d.*
Para cada bridge RSJ destino *d,* de los que se mantengan rutas, se guarda la siguiente información:
-Distancias (coste) mínimas y máxima estimadas.
-Puerto de salida hacia el bridge destino.
-Bridge RSJ siguiente salto hacia bridge RSJ destino.
-Coordenada RSTAA del bridge destino.

Coordenada RSJ:
Formato de la coordenada RSJ : De 64 bits de longitud: 3 bit (nivel) - 10 bit primera coordenada- 2ª-3ª……- 6ª coordenada. Formato: *nnn-1111111111-2222222222-3333333333-4444444444-5555555555-6666666666.*

TVD. Tabla de vectores distancia entre bridges AB.

Incluye distancias estimadas máximas y mínimas entre bridges RSJ, puerto de salida hacia bridge RSJ destino, ID del bridge RSJ de siguiente salto para llegar a destino, coordenadas RSTAA del bridge destino (del tipo p.q.n ...). La tabla contiene:
MAC bridge destino, Coordenada RSTAA bridge destino, dist. Mínima a b. Destino, dist. Max., puerto de salida para destino, MAC del siguiente RSJ (next hop), Relación y profundidad (antecesor, sucesor y profundidad relativa, ninguna relación).

Base de Datos de Reenvío (BDR). *Forwarding Database (FD).* Es la tabla estándar del AB como bridge. Incluye los sistemas finales de la tabla *TSF* y todas las direcciones MAC aprendidas por el bridge y que no han caducado. Las direcciones expiran por defecto a los 300 segundos.
*Tuplas: (ARB(MAC), host (MAC), puerto del bridge). Doble MAC learning.** Para resolver ARB destino.*
Las direcciones MAC de los ARBs en las tablas Host-ARB caducan cada 300 segundos si no se refrescan.

5.4.3 Servidores ARP

Los servidores ARP/AB mantienen la siguiente información como mínimo:
Tabla de Sistemas Finales: TSF. Por cada entrada de host: dirección IP, dirección MAC, coordenada RSTAA más próxima al host, AB asociado, vencimiento. La tabla se accede por dirección IP.

5.5 Tráfico Multicast

En esta sección se analiza la problemática del tráfico multicast descrita en el capítulo 3 y se describe la implementación propuesta para la optimización del tráfico multicast en en IPv4 e IPv6 así como posibles variantes.

En los capítulos 2 y 3 se ha tratado la problemática de los distintos componentes del tráfico multicast y el estado del arte de los protocolos de nivel 2 y 3 relacionados con dicho tráfico. El problema principal de multicast en capa 2 proviene del hecho de que los bridges inundan cuando la dirección destino MAC es una dirección de grupo multicast (GDA) debido a que, dado que las direcciones multicast no pueden utilizarse como direcciones origen en la trama Ethernet, los bridges no pueden aprender las direcciones multicast. En el caso hipotético de que pudieran aprender las direcciones multicast, sus especiales características requerirían una modalidad de aprendizaje de las direcciones distinta a las unicast, dado que las direcciones multicast pueden usarse en diferentes puertos a la vez.

Se definía el problema de la optimización del tráfico multicast básicamente como consistente en reducir la infraestructura de bridges y enlaces utilizada para el tráfico multicast, así como en minimizar la longitud de dichos caminos de difusión.

Existen dos tipos básicos de tráfico multicast: el utilizado por los protocolos de base, como los de encaminamiento para comunicarse entre nodos que ejecutan el mismo

protocolo y el tráfico en general de todo tipo, pero típicamente de formato multimedia, para usuarios que se suscriben a dichas fuentes de tráfico, tales como teleconferencia, trabajo en grupo, contenidos multimedia, etc.

Como se ha indicado, el tráfico multicast producido por los protocolos IP como OSPF, DHCP, etc, no es "podable", por lo que su optimización en principio no parece fácil ni siquiera recomendable o segura, dado que cualquier restricción en la difusión de los paquetes multicast hace posible que los paquetes no lleguen a dispositivos que deberían recibirlos. Por otro lado el ancho de banda ocupado por dicho tráfico es pequeño, dado que son paquetes de pequeño tamaño y se emiten con periodicidad baja.

Las opciones de diseño para el tráfico multicast en las redes campus plantean algunas decisiones alternativas: La primera consiste en si encapsular o no el trafico multicast para limitar su difusión y la segunda es el método de poda y suscripción a utilizar para limitar la difusión.

Respecto al posible encapsulado del tráfico multicast, existen las alternativas de encapsular en los ABs, en los ARBs, o no encapsular. El encapsulado en los AB no se justifica porque pueden estar situados de forma dispersa. El encapsulado en los ARBs no presenta este inconveniente, pero no ofrece ventajas claras. En ambos casos el encapsulado del tráfico multicast supondría complejidad añadida por la diferenciación de procesado con los bridges estándar. Por compatibilidad y coherencia, los ABs y ARBs deben soportar los protocolos estándar de optimización de multicast de capa 2 tales como inspección IGMP, además del protocolo específico elegido para optimizar. No vemos por el momento razones suficientes para encapsular el tráfico multicast en la red. Por tanto ni los AB ni los ARBs encapsularán el tráfico multicast, la dirección multicast destino será visible en las tramas (cabecera original inalterada) y los bridges que no tengan funcionalidad IGMP inundarán y tratarán el trafico multicast en nivel 2 como el de broadcast (difusión).

5.5.1 Tratamiento de tráficos multicast en redes campus

Se describe primero la propuesta para el tráfico multicast intensivo, a continuación la de protocolos de base de la red, mostrando su eficaz funcionamiento en el troncal y en la capa de acceso de la arquitectura propuesta.

5.5.1.1 Tráfico multicast de contenidos

Bajo el nombre de tráfico multicast de contenidos designamos al tráfico no ligado a protocolos ni a servicios de la red. El principal volumen de este tráfico vendrá generado por los contenidos multimedia y de intercambio de ficheros de todo tipo, y puede consumir, al contrario que el tráfico multicast de protocolos y servicios, un porcentaje significativo del ancho de banda de la LAN. Lo principales a optimizar en el tráfico multicast es un primer lugar suprimir en lo posible la difusión innecesaria del tráfico multicast (normalmente mediante poda (*pruning)),* en segundo lugar hacerlo llegar hasta los sistemas finales destinatarios utilizando la mínima infraestructura posible para transportarlo por caminos óptimos, y finalmente el balanceado de la carga de dicho tráfico para evitar congestiones. El origen del tráfico multicast puede ser en principio cualquier host.

Como se ha descrito en el capítulo 3, el problema del tratamiento multicast tiene dos partes: el tratamiento en capa 3, en el que intervienen encaminadores y sistemas finales y el tratamiento en capa 2 en el que el papel principal lo desempeñan los bridges. En entornos de multicast en modo denso (Dense Mode (DM)) el proceso consiste en crear los árboles multicast y podarlos. Para preservar la transparencia en multicast y la compatibilidad, ni los AB ni los ARBs pueden intervenir en el procesado de capa 3. El tratamiento de capa 2, para mantener la compatibilidad con los estándares de bridges, consiste, tanto en entornos en modo denso o disperso (Sparse Mode (SM)), en realizar inspección de los paquetes intercambiados entre sistemas finales y encaminadores para realizar la poda de la infraestructura innecesaria.

La arquitectura propuesta para unicast presenta ventajas para la difusión multicast. La construcción de árboles de difusión multicast en capa 2 sería normalmente compleja, pero las instancias de árboles múltiples del troncal enraizadas en cada ARB frontera, combinados con la poda, sirven como eficientes árboles de difusión multicast.

La solución tradicional para el tratamiento multicast, a fin de evitar la inundación (broadcast) es configurar las direcciones en los bridges como direcciones estáticas en los puertos donde están conectados los sistemas finales suscritos, pero requiere configuración manual.

El procesamiento actual en capa 2 del tráfico multicast de este tipo está soportado por la funcionalidad estándar de inspección IGMP (*IGMP snooping)* descrita en el capítulo 3. Esta funcionalidad permite la poda de las ramas no utilizadas, es decir las que no son atravesadas por los mensajes del protocolo IGMP intercambiados entre el router multicast y los sistemas finales que se suscriben al grupo multicast correspondiente. Requiere la inspección de todo el tráfico multicast entrante en el bridge, es una característica ya generalmente disponible en los bridges de gama media y alta. Esta es la solución recomendada por su compatibilidad y operatividad

Adicionalmente puede utilizarse el protocolo estándar de registro multicast GMRP. Este protocolo requiere menos procesamiento en los bridges dado que el registro en los grupos multicast se hace explícitamente al bridge. En cambio requiere que los sistemas finales lo soporten, lo que no es de aplicación general.

Algunos protocolos propietarios como CGMP (Cisco) están muy extendidos. Presentan la ventaja del menor procesado, dado que es el router el que informa a los bridges conectados a él de las suscripciones IGMP recibidas de los sistemas finales. Actualmente CGMP es reemplazado por inspección IGMP en los bridges con hardware avanzado.

En conclusión, se recomienda utilizar alternativamente los protocolos estándar de inspección IGMP y GMRP, quedando excluido CGMP por ser protocolo propietario (Cisco). Tanto los ABs como los ARBs (éstos de forma prioritaria) deben soportar GMRP e inspección IGMP.

5.5.1.2 Tráfico multicast de protocolos y servicios

Diversos protocolos de base de la red utilizan direcciones de grupo multicast para intercambiar sus PDUs. Entre ellos podemos mencionar el Protocolo de Árbol de

Expansión, los protocolos de encaminamiento como OSPF y otros. El objetivo de utilizar la dirección multicast es alcanzar a todos los dispositivos de un determinado tipo sin importar dónde se encuentren.

Además de estos protocolos básicos de la red, existen aplicaciones que anuncian sus servicios mediante direcciones multicast. De esta forma se evita tener que configurar en los clientes la dirección del servidor que provee dicho servicio. En ambos casos, el ancho de banda consumido por dichos protocolos y anuncios de servicios es bajo. El ancho de banda es bajo porque la periodicidad de los envíos está en el orden que va desde los 2 segundos de STP a decenas de segundos en el caso de los anuncios de servicio, siendo intermedio (10 segundos por defecto) para el protocolo OSPF. Como se ha descrito en el capítulo 3, dados los incrementos de las capacidades de Ethernet a 1 y 10 Gigabit, los porcentajes de tráfico de este tipo sobre la capacidad total se reducen hasta dos órdenes de magnitud sobre la cifra de 0,0003 % de la capacidad de una LAN de 100 Mbps consumida por las BPDUs del protocolo STP.

Como particularidad, mencionemos la existencia de periodos inferiores a 1 segundo en algún caso (por ejemplo los OSPF Fast Hello packets que pueden llegar a enviarse a una tasa de 20 paquetes por segundo), utilizados para acelerar la convergencia detectando la pérdida de un vecino en menos de un segundo. Estos son casos no estándar por el momento y que en todo caso suponen solamente un orden de magnitud más que los actuales de STP, por lo que la argumentación seguiría siendo válida.

La cuestión que se plantea es si procede el objetivo de reducir la difusión del tráfico de los protocolos de base. Actualmente la inspección IGMP no cubre el rango de direcciones de direcciones “well-known” multicast (224.0.0.1 a 224.0.0.255), por lo que no puede realizarse poda de la difusión de estos grupos multicast, poda que por otro lado supondría serios riesgos de mal funcionamiento. A esto se añade el escaso volumen de dicho tráfico por la baja periodicidad mencionado antes.

5.5.1.3 Multicast en troncal.

El protocolo AMSTP usado para el encaminamiento de tráfico unicast en el troncal es válido y óptimo también para el encaminamiento multicast. Efectivamente, AMSTP es óptimo para difusión multicast en el troncal porque crea árboles basados en nodo origen del tráfico multicast (*source based trees)*. En el troncal, los nodos origen de tráfico multicast son siempre los ARB frontera. AMSTP encamina en multicast utilizando árboles múltiples cuya raíz es cada uno de los bridges frontera. Por ello los caminos del tráfico multicast son mínimos de nodos origen a nodos destino a nivel troncal, satisfaciendo así uno de los criterios básicos de optimización del tráfico multicast: el de mínimo coste o mínimo retardo desde origen a destino. En la Figura 54 se muestran las instancias de árboles de expansión utilizadas por cada ARB.

Por el contrario, AMSTP no se adapta a otros criterios alternativos de optimización de los descritos en la definición del problema de multicast, tales como minimizar el coste total del árbol de difusión multicast *(Árbol de Steiner*) [Bla04] empleando nodos adicionales o criterios combinados de optimización. Pero se adecua perfectamente a la difusión multicast mediante árboles basados en fuente, lo que permite una optimización del tráfico multicast en el troncal usando el mismo encaminamiento que el tráfico unicast, extendido con las funciones de poda y unión propias del multicast.

La segunda ventaja de los árboles basados en origen es que reparten el tráfico, reduciendo el riesgo de congestión, que aumenta cuando se intenta minimizar el coste total del árbol de distribución multicast.

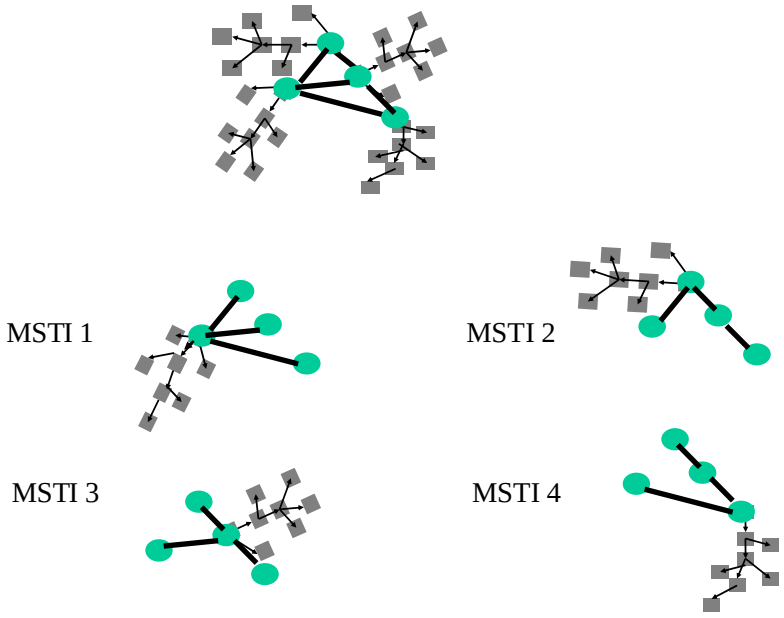


Figura 54. Instancias de árboles de difusión multicast en troncal

En la Figura 55 y la Figura 56 se muestra el funcionamiento de la inspección IGMP en los bridges. Los SF envían al encaminador multicast las suscripciones a una determinada dirección de grupo multicast, éstas son inspeccionadas por los bridges y procesadas entre ellos, de forma que las ramas que no tienen ningún SF suscrito a esta dirección, no recibirán difusión de las tramas correspondientes.

La poda en los árboles AMSTP se realiza como se ha descrito más arriba, mediante IGMP snooping. Los ARBs inspeccionan los paquetes de IGMP intercambiados entre los sistemas finales y los encaminadores multicast y actúan en consonancia podando las ramas no utilizadas e incluyendo las utilizadas según los contenidos observados de los paquetes de IGMP Query e IGMP Response.

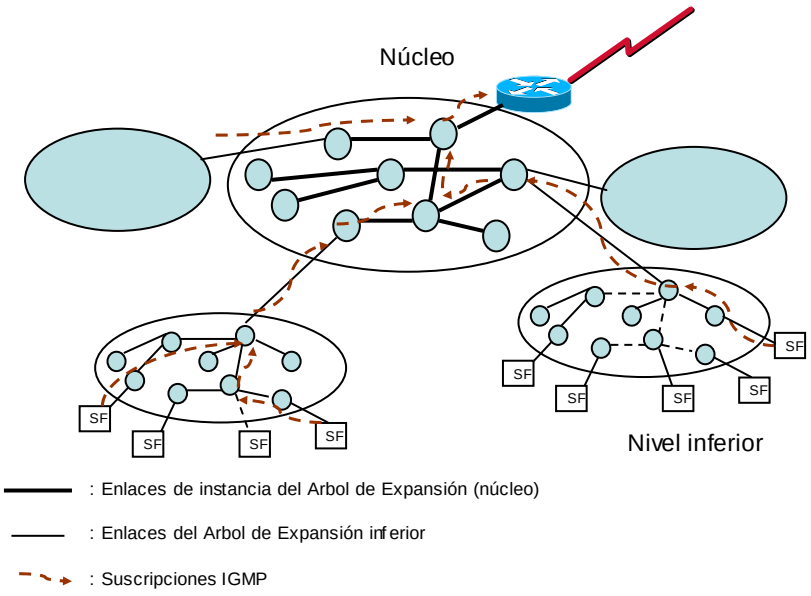


Figura 55. Suscripciones IGMP hacia el router multicast.

La utilización de árboles basados en origen en el troncal con AMSTP e IGMP snooping permite optimizar los flujos allí donde la infraestructura es más costosa. Dentro de la red campus, la mayor parte del tráfico multicast acabará llegando en algun momento a un ARB del troncal, desde donde será encaminado por el árbol enraizado en ese ARB hacia otros ARBs, las ramas que llegan a ARBs que tengan destinatarios en su área por haberse suscrito.

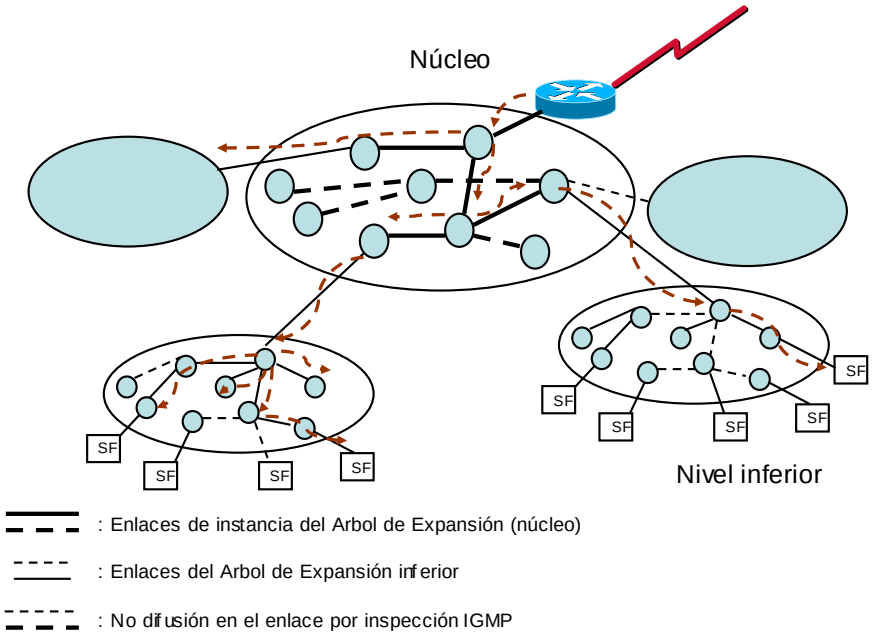


Figura 56. Difusión de tráfico multicast resultante con inspección IGMP

El tráfico multicast de tipo multimedia, de gran ancho de banda, presenta una alta variabilidad y es poco predecible, por lo que la matriz de tráfico inicialmente prevista en el diseño del troncal de la red puede cambiar sustancialmente. Esto aumenta la necesidad de optimizar la infraestructura para multicast y le añade el requisito de que esta optimización sea además lo más independiente posible de la matriz de distribución tráfico en la red campus. Nuestra propuesta para el troncal, sin llegar a la independencia de la matriz de tráfico que obtienen las redes Valiant de conectividad total mediante reparto de tráfico [Valiant], presenta características intrínsecas de reparto de tráfico en el troncal, aunque en caso de saturación del camino entre dos bridges no reparte el tráfico desviándolo por rutas alternativas. Las modificaciones del protocolo AMSTP para incluir comportamientos cercanos a los de las redes Valiant no parece excesivamente compleja si el ARB al percibir congestión en una ruta, desvía parte del tráfico a un ARB muy cercano en coste para que éste lo reenvíe al ARB destino. El comportamiento a este respecto del ARB intermedio sería un caso particular del reenvío, detectado por el ARB intermedio por ser él parte del camino entre las MAC origen y destino de la trama recibida, en otras palabras, no pertenecer al camino de árbol entre ARB origen y destino.

5.5.1.4 Multicast en nivel inferior

El tráfico multicast recorrerá normalmente los dos niveles de la red considerados. En el caso más general, la fuente multicast será un router multicast, situado en una de las áreas conectadas a un ARB. Veamos el funcionamiento en los niveles inferiores de la red. Desde el router multicast, el tráfico multicast será enviado, por los bridges del área solamente, mediante inundación y poda IGMP, hasta los sistemas finales que estén suscritos. Si hay suscripciones en otras áreas, se propagarán por el troncal mediante inspección IGMP de los mensajes efectuada por los ARBs del trayecto. Por lo tanto el camino de difusión multicast es el mostrado en la Figura 56.

La difusión multicast en el nivel inferior hasta el ARB de entrada en el troncal se realiza mediante RSTP por el árbol de expansión. La razón de ello es doble: por un lado el protocolo STAR no contempla la construcción de árboles para difusión multicast aunque sería susceptible de ser extendido para crear árboles de distribución multicast desde la fuente; por otra los caminos creados con RSTP en las redes inferiores son óptimos en todas las redes receptoras y cercanos al óptimo en la red emisora, debido a que los ARBs son raíz del árbol de expansión de la red. En [STAR] se indica que es posible montar árboles de multicast mínimos basándose en STAR. Adicionalmente pueden asignarse caminos más o menos óptimos en función de la prioridad mediante el protocolo PEARL. Pero consideramos que la efectividad sería muy baja para la complejidad añadida excepto en situaciones excepcionales en que el trafico multicast estuviera confinado en cada área inferior y no circulara de forma significativa por el troncal.

5.5.2 Tratamiento de multicast en ABs y ARBs

El resultado del análisis efectuado de las posibilidades de utilizar encapsulado en ABs y ARBs para reducir la difusión del tráfico de los protocolos de base (*well known multicast addresses*) es negativo, como se detalla en apartados posteriores.

Por sencillez y compatibilidad con el modelo estándar de multicast existente, la solución adoptada consiste en la difusión sin restricciones por los ABs y ARBs de las tramas multicast de los protocolos de base. El procesamiento de una petición de un sistema final será entonces el siguiente: el AB agente del sistema final origen de la trama multicast reenvía la trama por el Árbol de Expansión inundando (no hay poda para estos protocolos dado que IGMP no lo soporta) y ésta llega al ARB. El ARB reenvía la trama por el troncal sin modificarla hacia los demás ARBs. Cada ARB redifunde en su árbol. Los servidores alcanzados contestan a la petición si les corresponde. La contestación recorre un camino similar, pero generalmente en unicast, dependiendo de los parámetros del servicio correspondiente. Como ejemplo, en el caso de DHCP, el SF recibe diversas ofertas de diferentes servidores DHCP y elige una.

La difusión multicast en el troncal utiliza el árbol de expansión del que el ARB frontera origen es raíz porque el encaminamiento (la pertenencia de la trama al árbol) se determina en multicast por dirección MAC origen de la trama (del puerto del ARB origen de entrada al troncal). Los servidores DHCP pueden estar situados en cualquier parte de la red campus.

5.5.3 Situación actual de DHCP

Como problema de referencia del tráfico multicast de los protocolos de base, se describe de forma resumida el funcionamiento actual del protocolo DHCP en IPv4 e IPv6.

DHCP

El diálogo estándar del protocolo DHCP se muestra en la Figura 57 y es como sigue: el cliente del SF envía la solicitud *DHCP discover* a la dirección multicast DHCP. En IP v4 se utilizan la dirección multicast: 224.0.0.12 DHCP Server/Relay Agents [IANA04]. Debe existir por cada LAN un servidor o *relay* DHCP. En IPv6 las direcciones utilizadas son: *FF02::1:2 All DHCP agents link-local* y *FF05::1:3 All DHCP servers site-local* (RFC 2275). Las respuestas del servidor DHCP dependen de cómo haya puesto el cliente el bit de broadcast (*flag*) en la petición. Si el cliente DHCP no soporta unicast IP cuando aun no ha recibido una dirección IP, el flag de broadcast debe haber sido puesto a uno para que el servidor sepa que debe contestar en broadcast. Si el bit está a 1 se contesta a la dirección IP broadcast (preferentemente 0xffffffff) y a la dirección de broadcast de capa 2 según la RFC 2131 [Dro97]. Se utiliza el puerto 67 en la dirección de cliente a servidor y el puerto 68 en la opuesta.

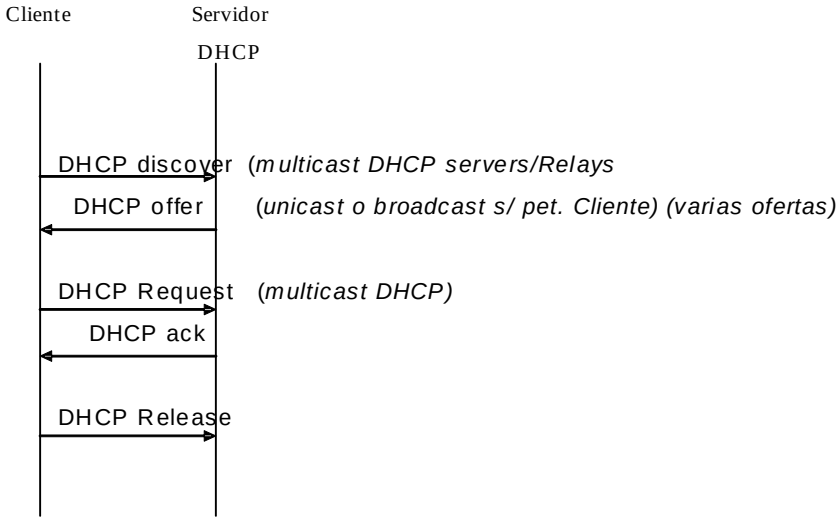


Figura 57. Diálogo DHCP

La secuencia es:

- El mensaje *DHCP DISCOVER* se envía a la dirección multicast.
- El servidor contesta con el mensaje *DHCPOFFER* en unicast salvo si el cliente DHCP no soporta unicast IP sin dirección IP asignada (bit activado en *discover*).
- El cliente puede recibir varios *DHCPOFFER*, elige uno y contesta con multicast al servidor con la mejor oferta, aceptándola. (*DHCPREQUEST*).
- El servidor envia en unicast ó broadcast el *DHCPACK* para confirmar.
- Finalmente el cliente del SF envía en unicast el mensaje *DHCPRelease*, para dejar disponible la dirección.

5.5.4 Requisitos del servidor agrupado

Se analiza aquí la factibilidad y conveniencia de un tratamiento agrupado de los protocolos y servicios multicast mediante uno o varios servidores agrupados que atiendan a todos estos servicios, organizados de forma jerárquica de forma que escale a un servidor superior los que no puedan resolver directamente. Este análisis se realiza independientemente de la implementación propuesta anteriormente para los servidores ARP/AB.

El esquema de funcionamiento con un servidor agrupado (*M-server*) es el siguiente: el AB agente del host recibe la trama enviada por el sistema final y destinada a un servicio que utilice broadcast o multicast, la encapsula con dirección destino la dirección unicast del M-server y la reenvía. El M-server la desencapsula, inspecciona e intenta resolver la petición, respondiendo al AB en unicast (de forma similar a la descrita para ARP). Si el M-server no puede resolver la petición, se escala hacia arriba hasta un servidor central y si dicho servidor tampoco puede resolverlo, se redirigirá la petición

encapsulada a la dirección multicast. Los requisitos funcionales de un posible servidor agrupado de este tipo serían :

- Discernimiento de direcciones. Debe reconocer las direcciones multicast Ethernet e IP en IPv4 y sus homólogas en IP versión 6.
  - En IP versión 4 el discernimiento de direcciones multicast es como sigue:
    - Rango de 224.0.0.1 a 224.0.0.255: *Well-known protocols*
    - Rango de 224.0.1.0 a 238.255.255.255: direcciones globales multicast
    - Rango de 239.0.0.0 a 239.255.255.255 : Direcciones locales multicast . Son similares a las direcciones privadas unicast, de ambito local o de organización.
  - Adicionalmente, en IP versión 4, para detectar las peticiones ARP, debe discernir tramas con destino dirección broadcast de capa 2 y el identificador de protocolo ARP en la trama.
  - En IP versión 6, las direcciones multicast se distinguen por el prefijo FF00/8, ocupando el identificador de grupo 112 bits, de los que los menos significativos coinciden con los valores de los grupos IP version 4, si existe protocolo equivalente en IP v6. La dirección multicast de todos los sistemas finales link-local es FF02::1
  - Descubrimiento mutuo entre los servidores agrupados, de igual o diferente nivel existentes en la subred. Opcionalmente: reparto de carga entre servidores de igual nivel. Delimitar su dominio o área de competencia.
- Discernir protocolo de la trama. El protocolo en multicast se identifica por la dirección de grupo multicast. Para más detalles del protocolo es necesario inspeccionar el formato del protocolo y ver versión del protocolo, tipo de mensaje, etc.
- Procesar el mensaje según servicio: ARP, etc.
  - Escalar los mensajes cuyos servicios no pueda resolver.
  - Si es el último servidor, difundir en toda la red hasta encontrar el servidor buscado.

Inspección del paquete IP.

Los campos a procesar son:

- Ethertype para discriminar IP v6 e IP v4.
- Dirección broadcast Ethernet mas protocolo ARP para discriminar ARPs.
- Dirección IP para identificar dirección de grupo multicast.
  - En IPv4 e IPv6 Por las *wellknown multicast addresses* en capa 2 directamente se puede identificar el servicio/destinatarios:
  - FF01:2 all DHCP agents
  - FF00:2 all routers
  - FF00:1 all nodes.
  - FF02:0:0:0:0:1:FFXX:XXXX *solicited node address*

Ejemplo DHCP con M-server

Como ejemplo de implementación con M-server describimos la de DHCP. La petición DHCP (DHCP DISCOVER) de un sistema final a la dirección de capa 2 de multicast DHCP es interceptada por su bridge agente AB, el cual la reenvía encapsulada en unicast al M-server. La dirección del M-server es conocida por los ABs porque los servidores M-server se anuncian en la dirección multicast *todos los bridges AB y ARB* con su dirección unicast. DHCP puede implementarse como un servidor independiente o de forma conjunta con otros servidores. Si se hace conjunta, se realiza la función DHCP

server localmente encapsulando la contestación al AB origen. Lo mismo es aplicable a la función DHCP relay.

5.5.4.1 Combinación de DHCP y ARP

Se hacen a continuación algunas consideraciones sobre la combinación de las funciones DHCP y ARP en un servidor único.

DHCP es parte de la autoconfiguración y registra las direcciones asignadas, por lo que puede contestar ARPs con certeza y autenticidad. En el fichero de lease (cesión) de direcciones del servidor DHCP se registra el nombre del host, la dirección IP, instantes de asignación y finalización de la asignación de la dirección IP y dirección MAC del host. De esta forma el servidor DHCP puede hacer simultáneamente la labor (inversa) de servidor ARP, pero hay cuestiones a tener en cuenta:

- En DHCP pueden contestar varios servidores a una sola petición, el host elige una de las direcciones ofrecidas.
- El servidor DHCP puede o no usarse, el servidor ARP será siempre utilizado.
- DHCP puede también usarse para asignar direcciones estáticas (fijas).
- Los criterios de direcciones IP de los servidores ARP deben ser consistentes con DHCP. Este no es el caso del reparto de carga en servidores ARP mediante hash(IP) porque direcciones consecutivas de IP darán distintos hash como resultado y distintos servidores encargados de ellas. Parece más compatible con DHCP el registro ARP por rangos de direcciones IP que el de reparto de carga por hash(IP).

Desde un punto de vista práctico, si el servidor conjunto M-Server agrupa las funciones de DHCP y ARP server, la combinación de funciones puede emplearse para controlar el acceso a la red y reforzar la seguridad ante suplantación de direcciones MAC a autenticación y establecer medidas de seguridad. Si se utiliza un servidor separado DHCP, éste no interviene en el registro del sistema final en el servidor ARP, con el resultado de menor seguridad pero mayor compatibilidad y flexibilidad.

5.5.5 Procesado de paquetes en ABs y ARBs

Se describe ahora el procesado de los paquetes multicast en los bridges de tipo AB y ARB descritos en esta Tesis de acuerdo al tratamiento multicast elegido.

Inspección IGMP. Al recibir un paquete *IGMP report*, el AB debe añadir a la tabla (CAM) de datos de reenvío (FD) el sistema final suscrito y puerto del bridge por el que se ha recibido para incluirlo en las difusiones futuras. Al recibir un paquete *IGMP leave*, el AB realiza lo contrario, elimina la dirección MAC de la tabla.

En IP versión 6, IGMP está incorporado dentro de ICMP. Se denomina Multicast Listener Discovery (MLD) y utiliza tres tipos de mensajes ICMPv6: *MLD query*, *membership* y *done*. El procesado será el correspondiente de forma similar a IP versión 4 respectivamente. IGMP permite limitar la restricción del tráfico multicast sin necesidad de software específico en los sistemas finales.

GMRP. GMRP se implementa en el software del sistema final y en el del bridge y en el sistema final se utiliza normalmente con IGMP. El sistema final genera paquetes de capa dos GMRP que corresponden a los paquetes de capa tres IGMP generados por el sistema final. El AB recibe ambos y utiliza los paquetes GMRP para limitar la difusión del tráfico multicast al host suscrito.

El procesado de multicast en los ARBs es en principio igual que el de los ABs.

El tráfico ARP se procesa según se ha descrito en el epígrafe correspondiente. El procesado del tráfico DHCP se realiza de la forma estándar, sin encapsular, mediante servidores DHCP.

5.6 Aspectos de seguridad

Aunque los aspectos de seguridad de las redes campus no son objetivo específico de esta Tesis, no puede pasarse por alto la problemática de los dominios conmutados en cuanto a la seguridad.

A medida que las redes Ethernet se han multiplicado y su tamaño medio ha crecido de forma considerable, la conciencia de la problemática de seguridad en capa 2, antes relativamente escasa, ha aumentado de forma notable. Muchos protocolos como ARP fueron diseñados para optimizar desde el punto de vista de facilitar el funcionamiento de la red en cualquier condición, porque no existía conciencia de los riesgos, mientras que hoy la seguridad tiene una importancia fundamental.

Los ataques en capa dos son importantes por varias razones: la primera es que a menudo los sistemas de detección de intrusiones no monitorizan ataques de capa dos, la segunda es que la extensión de los dominios de capa dos es cada vez más grande y la tercera es que ya hay proveedores de acceso a Internet que ofrecen acceso en capa dos, con lo los riesgos de ataques de atacantes internos son mucho mayores y el concepto de *atacante interno* adquiere un significado distinto.

El planteamiento de seguridad empleado en [Rbridge] nos parece válido también para el ámbito de esta Tesis, aunque algo escaso. Lo podemos formular de la forma siguiente: no siendo objetivo primario de los nuevos dispositivos resolver el tema de la seguridad en las redes conmutadas, al menos no deben agravarlo ni perturbar los mecanismos de seguridad que se habiliten en los bridges convencionales. En las redes campus que contemplamos, los aspectos de seguridad tienen la misma importancia que para las redes campus actuales, aunque incrementada por su mayor tamaño. Este superior tamaño incrementa tanto los riesgos como los posibles daños. Así pues los requisitos mínimos de seguridad deben ser tales que al menos las propuestas no reduzcan la seguridad, siendo muy deseable que la aumenten. Nuestra propuesta de usar servidores para ARB se orienta en esa línea de aumento de la seguridad.

Como en todas las redes, aparece la disyuntiva entre seguridad por un lado y simplicidad y disponibilidad de la red por otro, lo que va en detrimento de la autoconfiguración o al menos de la simplicidad de configuración, objetivo explícito de

estas redes. Dado que la autoconfiguración es un requisito muy deseado, deberá decidirse en cada caso el compromiso entre seguridad y sencillez de configuración y explotación.

Uno de los aspectos clave para la fiabilidad y seguridad de las redes campus es la compartimentación. En las redes actuales dicha compartimentación se consigue mediante subredes IP distintas, interconectadas mediante encaminadores o mediante lo que se suele denominar como conmutadores de capa 3 o conmutadores multicapa. Esta compartimentación limita los dominios de difusión y a la vez, en caso de fallo, limita las áreas afectadas a una, en vez de afectar a todo el dominio de difusión. Estos fallos pueden deberse a una tormenta de tramas o a un error de configuración. En nuestra propuesta la compartimentación de las redes se realiza mediante los ARBs, que separan dominios de conmutación permitiendo su aislamiento en caso de fallo. La limitación de los dominios de difusión de las tramas se realiza con los servidores ARP principalmente.

5.6.1 Análisis de vulnerabilidades

En este epígrafe se analizan las posibles vulnerabilidades que presenta la propuesta, y se comparan con las habituales de las redes de capas 2 y 3. Las vulnerabilidades principales en capa 2 son las que se describen en los siguientes epígrafes.

5.6.1.1 Suplantación de MACs y vulnerabilidades de ARP

Un ataque de suplantación de MAC se produce cuando la dirección MAC origen se falsifica, reemplazándola por la de otros equipos detectados en la red por el suplantador. Un ataque de suplantación de MACs puede atraer todo el tráfico del conmutador al segmento atacado para denegación de servicio en la LAN.

Por otra parte ARP es un protocolo fácilmente atacable. Los principales ataques a ARP son: Corrupción de cachés ARP (basta enviar falsos paquetes de *ARP response* no solicitados para que la MAC transportada sea incorporada a la caché ARP del host). Mediante la suplantación de ARP se hacen posibles los ataques *Man-in-the-Middle (MiM).* Las vulnerabilidades de ARP han sido descritas como el “eslabón débil” entre las capas 2 y 3 [Alt+04]. El problema básico de la vulnerabilidad de ARP es la inexistencia de ligazón (*binding*) segura entre las direcciones IP y MAC. Las soluciones propuestas actualmente incluyen:

- DHCP seguro con diversas propuestas de DHCP con autenticación (IETF WG DHCP rfc3118) incluyendo en algún caso un servidor de confianza y la utilización de mecanismos de seguridad en capa 2. Además es necesaria la autenticación de las tramas que transportan mensajes ARP.
- Nueva arquitectura de nombrado. Host Identity Protocol (HIP) es un ejemplo de ello. Puede resolver los ataques de suplantación de ARP si todas las transacciones se realizan de forma segura, pero no los ataques al conmutador para desviar el tráfico por suplantación de MAC o por saturación de la tabla del conmutador para que pase a modo difusión.
- En el IEEE está en proceso de aprobación (septiembre 2005) la preparación de un nuevo estándar IEEE P802.AR Secure Device ID que posibilitará ARP seguro.
- Asimismo en el seno del IEEE, el grupo 802.1AE MAC Security Task Group prepara el estándar P802.1AE para garantizar los datos de usuario, la integridad y

autenticidad del origen de las tramas. Pero se concentra en garantizar la integridad punto a punto, no la integridad global de las tramas.

La arquitectura propuesta permite implementar fácilmente mecanismos de autenticación detección en el AB o en el ARB

5.6.1.2 Ataques al protocolo de árbol de expansión

Son posibles los ataques a los bridges mediante BPDUs del Protocolo de Árbol de Expansión falsas y otros medios tanto en el caso del protocolo STP como de su sucesor RSTP [AM03][AVM02][Mar03a][Mar03b]. Esta vulnerabilidad se debe principalmente a que el protocolo de Árbol de Expansión no autentica el bridge origen de las BPDUs, por lo que un bridge puede difundir información falsa. Existen diferentes formas de ataque:

- Reconfiguraciones continuas: se consiguen ataques de Denegación de Servicio forzando la reconfiguración continua del Árbol de Expansión mediante BPDUs falsas (muy ventajosas de coste para el destinatario) que fuerzan la elección bien sea de bridge raíz, bridge designado o puerto raíz.
- Pueden inducirse bucles en la red mediante ataques cíclicos.
- Denegación local de servicio en un bridge para impedir el acceso de un determinado host a un servidor y reemplazarlo por un servidor falso. El atacante, conectado al mismo bridge que el host, envía BPDUs falsas para conseguir ser elegido bridge designado del bridge al que están conectados el atacante y el host hasta conseguir deshabilitar el enlace entre dicho bridge y el bridge del servidor auténtico.
- Saturación de las tablas del bridge. Los bridges en caso de saturación de tráfico pasan a operar como concentradores para evitar la pérdida de tramas, por lo que los ataques que saturan su capacidad fuerzan el fallo. El modo concentrador permite también la captura de datos sensibles de todo tipo de protocolos, no solamente IP, por operar en capa dos.

Este tipo de ataques es posible tanto en la red troncal (protocolos AMSTP), como en la red inferior (protocolo RSTP). La medida más segura consiste en habilitar 802.1X, lo que presupone conexiones punto a punto dedicadas y control de acceso a los puertos mediante autenticación ante servidor Radius. IEEE 802.1X no protege frente a ataques de bridges autorizados, por lo que si un ARB autorizado se comporta como atacante puede bloquear la red.

5.6.1.3 Ataques a los protocolos de encaminamiento.

Si no hay autenticación entre ABs y ARBs, es factible el ataque a los protocolos de encaminamiento, difundiendo informacion falsa tanto de ABs unidos a los ARBs como de distancias entre ABs. La solución consiste en usar autenticación entre los AB.

5.6.2 Mecanismos de seguridad

Son numerosos los mecanismos de seguridad propuestos e implementados para bridges. No es posible extenderse en las implicaciones, ventajas, inconvenientes y aplicabilidad de cada uno a las redes campus propuestas, pero enumeraremos los que consideramos más significativos para redes conmutadas actuales y los adicionales que

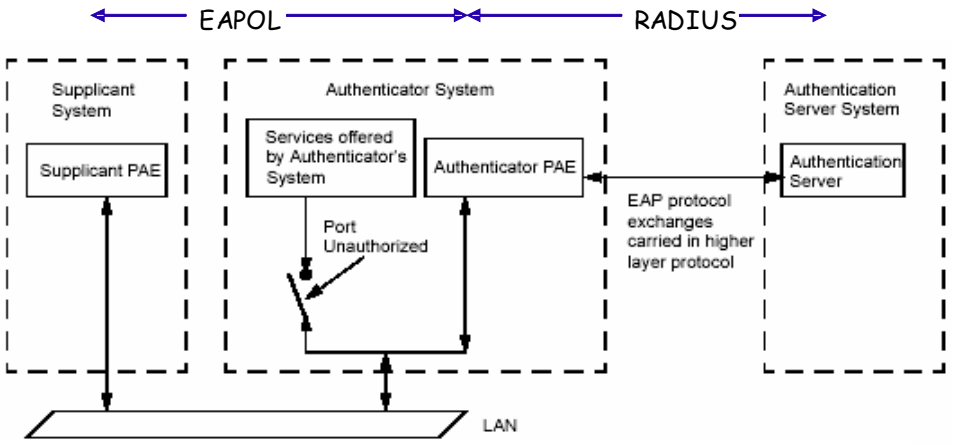
contemplamos para las redes propuestas, materializados en los protocolos y procesamiento en los ABs y ARBs.

5.6.2.1 Mecanismos de seguridad en capa 2.

Los principales mecanismos de seguridad que se proponen para hacer frente a las vulnerabilidades mencionadas son:

Inspección DHCP en el bridge de entrada. Cuando se utiliza DHCP para el direccionamiento, la inspección en el bridge de entrada de la correcta correspondencia entre direcciones MAC e IP permite distinguir a los paquetes ARP falsos. Este mecanismo elimina algunas importantes vulnerabilidades basadas en suplantación de MAC, pero requiere el uso de DHCP, lo que no constituye un inconveniente necesariamente, dado que permite configurar las direcciones IP tanto estáticas dinámicamente.

Control de Acceso por Puerto. IEEE 802.1X. El estándar IEEE 802.1X permite, en enlaces dedicados, el control de acceso basado en puerto. En la Figura 58 se muestra el esquema básico de autenticación y control de acceso. Mediante el mecanismo de control de acceso basado en puerto, el bridge no permite el encaminamiento de tramas que no sean las estrictamente necesarias para autenticar el puerto del host conectado ante el servidor de autenticación.



PAE= Port Access Entity : Suplicante , Autenticador , Servidor de Autenticación.

Figura 58. Entidades en protocolo IEEE 802.1X

La secuencia de mensajes de autenticación en 802.1X se muestra en la Figura 59. Las entidades básicas que actúan son el Suplicante, que solicita el acceso para enviar tramas, el Autenticador, normalmente el bridge al que esta conectado el suplicante, y el Servidor de Autenticación, que colabora en la autenticación con el Autenticador. De esta forma puede asegurarse que los sistemas finales conectados son legales. Evidentemente un sistema final legal puede comportarse de forma malintencionada, pero un usuario legal está mucho más controlado que un intruso.

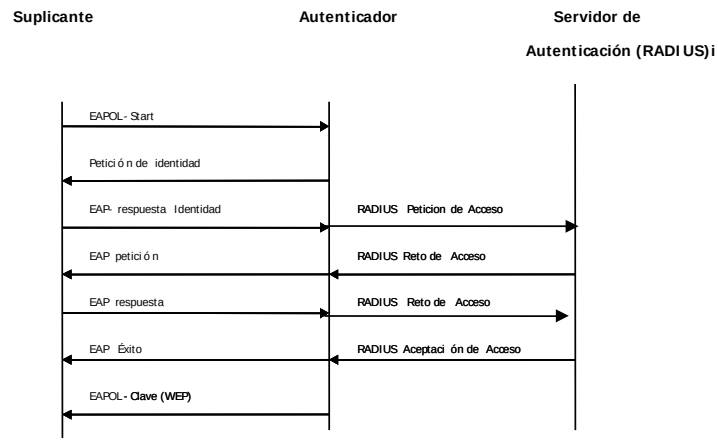


Figura 59. Secuencia de acceso controlado por puerto IEEE 802.1X

5.6.2.2 Mecanismos adicionales

Además de los mecanismos de seguridad descritos, se pueden implementar diversas medidas en los AB/ARB para el control de ataques:

- Disparo de alarmas de intrusión ante excesivo número de ARPs y MACs nuevas.
- Limitación del servicio de paquetes ARP con limitación de respuesta por host, por unidad de tiempo, etc. Este tipo de medidas son mucho más efectivas implementadas en los AB que en los servidores ARP.
- Limitación del número de MACs aprendidas por puerto.
- Mecanismos de *Root Guard* y *BPDU Guard.* [Cisco]

Como conclusión, vemos que la arquitectura propuesta, basada en el procesado por los ABs y la asignación de cada sistema final a un Bridge Agente (AB), permite distribuir las medidas de seguridad entre los ABs y prevenir los ataques en su origen. En el escenario simplificado, al no utilizarse ABs, las medidas de seguridad recaen completamente en los ARBs y en los servidores ARP.

5.7 Conclusiones

Se han descrito los protocolos nuevos, derivados o extendidos, propuestos para el núcleo y troncal para unicast y el procesado de tráfico para multicast, que mediante protocolos estándar de inspección IGMP permite una supresión de tráfico muy eficiente en el núcleo y caminos mínimos de multicast, como los de unicast. Se han descrito las

tablas y estructuras de datos requeridas por los protocolos en los dispositivos. Se ha analizado la posibilidad de servidores agrupados, sin hallar especiales ventajas, salvo en algunos escenarios específicos la agrupación de servidores DHCP y ARP, razón insuficiente para recomendarlos. Se han descrito las principales vulnerabilidades de las redes campus descritas y las principales medidas preventivas. La red propuesta cumple los objetivos de seguridad equivalente y no presenta mayor vulnerabilidad que las redes actuales. Por el contrario, el paso obligado para cada sistema final por un determinado bridge AB (en el escenario completo) ó un ARB (en el simplificado) permite mejorar los mecanismos de defensa, de detección de ataques y facilitar la autenticación y la localización del atacante.

Capítulo 6

Análisis de prestaciones

En este capítulo se realiza la evaluación de la propuesta respecto a los requisitos, mostrando el grado de autoconfigurabilidad de la red obtenida, verificando las prestaciones de la arquitectura y de los protocolos propuestos, comparándolas con las del Protocolo de Árbol de Expansión, con las de Rbridges y de otros protocolos de caminos mínimos con conectividad completa, así como otros aspectos relevantes. Se analizan la complejidad de mensajes y las necesidades de almacenamiento de los protocolos propuestos.

6.1 Metodología

Como se ha descrito en el capítulo precedente, la topología jerárquica de la red propuesta consta de dos capas: una capa con funciones de troncal y distribución y otra capa con funciones de acceso, con protocolos relativamente independientes para ambas. En la capa de acceso existen varios protocolos de uso alternativo. Esto hace recomendable una evaluación separada del rendimiento de cada capa de red para obtener el rendimiento combinado después. Se expone primeramente la evaluación de prestaciones del Núcleo o red troncal y después las del nivel de Acceso. En cada una de las dos redes se describen los recursos necesarios en lo que respecta a necesidades de almacenamiento y proceso y la complejidad de mensajes utilizados por los protocolos propuestos. Se evalúa para el núcleo el rendimiento obtenible en redes tanto de topología aleatoria como regular, dada la adecuación de las topologías regulares de alta conectividad para el núcleo de redes campus. En el nivel de acceso se evalúa el

rendimiento en topologías aleatorias. Finalmente se ha simulado el funcionamiento del mecanismo de doble bridge raíz propuesto para el funcionamiento de ARBs por parejas en modo activo-activo.

Los protocolos utilizados en el escenario simplificado son AMTSP en la capa troncal y RSTP en la capa de acceso. Opcionalmente, en la capa de acceso pueden utilizarse los protocolos RSTAR ó RSJ, empleándose RSTAR cuando se prefiera la máxima compatibilidad con los bridges estándar y RSJ cuando se desee una utilización más eficiente de la infraestructura y mayor escalabilidad. Alternativamente, RSJ es susceptible de utilización como protocolo único en redes campus de tamaño medio. Con la estandarización de RSTP, la factibilidad de redes completas de tamaño medio con RSJ como único protocolo parece menos lejana dada la posibilidad de migración software software de los bridges de RSTP a RSJ.

La evaluación del rendimiento de una red campus de gran tamaño plantea problemas específicos: por un lado, por el propio tamaño de la red en cuanto al número de nodos, la multiplicidad de combinaciones; y por otro por las múltiples alternativas de valoración y combinaciones de redes posibles.

La independencia entre los protocolos de troncal y de acceso permite una evaluación separada de los rendimientos de ambos protocolos. En redes jerárquicas en dos niveles, como la propuesta, la mejora efectiva obtenida en una red concreta dependerá de los perfiles de tráfico a cursar, en concreto de la proporción entre tráfico local (denominado *intra-área)* y tráfico vía troncal (también denominado *inter-área).* Dado el predominio actual del modelo cliente-servidor y la ubicación de las granjas de servidores (normalmente cercanas a la red troncal) es razonable esperar que predomine en la mayoría de los casos el tráfico área-troncal, aunque es imposible predecir con seguridad la evolución de la matriz de tráfico, sobre todo si el modelo *Peer to Peer (P2P)* llega a ser dominante en las redes campus en el futuro.

Por otra parte, dado que se trata de redes campus con referentes actuales concretos y con problemática concreta y requisitos ligados a costes de equipos, es conveniente no dejar de lado la problemática práctica de diseño y dimensionado de estas redes, sus modelos de costes y la tipología de los equipos utilizados actualmente. Este aspecto lo tratamos de forma específica al final del presente capitulo, donde concretamos algunos posibles casos de aplicación de la red propuesta.

De acuerdo con lo expuesto, se describe primeramente la evaluación del nivel troncal y del protocolo AMSTP, tratando después la capa de Acceso y los protocolos RSTP, RSTAR y RSJ, el rendimiento conjunto de ambas capas, los aspectos de diseño y dimensionado de redes, las necesidades de los servidores ARP y unas consideraciones finales.

Se utilizan tres métricas básicas para medir la eficacia en árboles de expansión. La primera es el porcentaje de utilización de los enlaces de la red: $U= 100* Ea / Et$, donde *Et* es el número total de enlaces y *Ea* es el número de enlaces activos. La segunda y tercera métricas empleadas consisten en la comparación de costes de caminos en número de saltos, camino más largo y camino medio respecto al camino mínimo entre nodos respectivamente.

6.2 Nivel Troncal. Prestaciones del Protocolo AMSTP.

El protocolo del nivel troncal, AMSTP, se basa en árboles múltiples de expansión. Como se ha indicado en su descripción, y a diferencia de MSTP, AMSTP es un protocolo de bridging de caminos mínimos (SPB en terminología IEEE), lo que hace equivalente su rendimiento al encaminamiento Djikstra y permite la utilización del simulador ns2 [NS2] para la evaluación de prestaciones. Al contrario de lo que sucede con RSTP, no están disponibles simuladores de los protocolos de caminos múltiples como MSTP, lo que dificulta la implementación de simuladores de AMSTP.

Se describen en primer lugar las necesidades del protocolo relativas a configuración, almacenamiento y complejidad computacional, y después los aspectos cuantitativos del mismo. Al tratarse de caminos mínimos (inmejorables), se trata principalmente obtener las longitudes de camino esperables en el núcleo, además de compararlas con el protocolo existente en capa 2 (RSTP), de manera complementaria.

6.2.1 Configuración de ARBs

En el epígrafe Bridge IDs de un ARB se han descrito las recomendaciones prácticas para asignación de Bridge IDs en los ARB a efectos de operación del protocolo AMSTP en el troncal y de RSTP, RSTAR o RSJ en la capa de acceso.

La configuración requerida por los bridges ARB para el protocolo AMSTP es la siguiente:

- Debe elegirse el ARB a actuar como bridge raíz y su bridge reserva del árbol RSTP principal del troncal y configurarlos respectivamente con el valor numérico más bajo de prioridad (prioridad más alta) de bridge ARB y el siguiente más bajo. El resto de ARBs pueden utilizar la prioridad por defecto. Equivale a la configuración estándar de árbol de expansión en una red.
- No se requiere configurar identidades (prioridades) de bridge en los demás bridges.
- No se requiere configurar VLANs en el núcleo.

La configuración requerida por los bridges ARB para los protocolos alternativos RSTP, RSTAR ó RSJ en la capa de Acceso es como sigue:

- La bridge ID no requiere configuración. Se recomienda emplear una segunda dirección MAC (programada de fábrica) para mayor claridad en la operación del árbol RSTP respecto al árbol RSTP de troncal. La prioridad se configura automáticamente al valor de máxima prioridad al operar como ARB. Esta prioridad puede modificarse automáticamente a una prioridad baja en caso de fallo de funcionamiento del ARB.

- Si la red de acceso tiene conexión a más de un ARB, para obtener redundancia en la conexión al troncal y mejorar la disponibilidad de la red, el primer bridge situado bajo el ARB debe implementar una variación sobre el protocolo RSTP en lo relativo a la construcción del árbol de expansión como se ha indicado en el epígrafe Modos de Funcionamiento de los ARBs en cuanto a la comparación de prioridad para elegir bridge raíz. Su prioridad por defecto debe ser más baja que la prioridad por defecto de los ARB para no ser nunca elegidos bridge raíz salvo en caso de fallo de todos los ARBs. Esto se logra haciendo que la prioridad por defecto de los ARB sea inferior a la prioridad por defecto de los bridges estándar.
- Si se utilizan VLANs, deben configurarse los conmutadores con la información de pertenencia a las respectivas VLANs.

6.2.2 Almacenamiento en ARBs

El almacenamiento requerido en los ARB se muestra a continuación. Cada ARB utiliza las siguientes estructuras de datos:

Tabla de Localización de ABs (TLA). Esta tabla contiene la lista que asocia los ABs con su ARB padre o cabecera del árbol. Es una lista de tuplas que contiene cada una: identidad_AB e identidad_ARB (direcciones MAC). Siendo M el número de ARBs, N el número de sistemas finales, A el número de ABs, el número de tuplas de la tabla es O(A), por tanto la longitud de la tabla TLA depende del número de ABs. Por razones de dimensionado el número de AB es en promedio de un orden de magnitud mayor que el de ARBs, siendo su máximo a todos los efectos 100*M, por lo tanto es O(M). Esta tabla no es necesaria en el escenario simplificado (AMSTP y RSTP).

Base de Datos de Reenvío (BDR). Es la caché MAC del ARB como bridge. Cada tupla contiene: *(ARB(MAC), puerto de salida del bridge)*. Como las tramas circulan encapsuladas, las direcciones MAC aprendidas por el ARB son las de los otros ARBs en los puertos troncales. Es *O(M)*. En los puertos del ARB que ejecutan 802.1D estándar las direcciones MAC aprendidas serán las de los bridges correspondientes.

Asociación de puerto a ARB. Esta tabla contiene las asociaciones de direcciones MAC de puerto a cada ARB. Cada *bridge_ID* de un ARB se asocia a la MAC del puerto designado de ese ARB raíz del que provienen las BPDUs AMSTP recibidas. Es *O(M)*.

6.2.2.1 Árbol RSTP.

La estructura de encaminamiento de RSTP utiliza lo que se denomina un *Vector de prioridad de puerto* que contiene:

- *bridge_id_raíz:* identidad del bridge raíz.

- *coste_a_raíz:* coste del camino hasta el bridge raíz.

- *bridge_id_designado:* identidad del bridge designado.

- *id_puerto_designado*: identidad del puerto designado (emisor de la BPDU).

- *id_puerto_receptor*: identidad del puerto designado (receptor de la BPDU).

Por cada puerto se utiliza una identidad de puerto designado para el envío y una identidad de puerto receptor para la recepción de las BPDUs. A diferencia de los protocolos de encaminamiento, el consumo de memoria es independiente del número de nodos M. Esta característica es propia de los árboles de expansión.

6.2.2.2 Árboles Múltiples (AMSTP)

Se describen ahora las necesidades de almacenamiento del procotolo AMSTP. Por cada instancia de árbol (una instancia por cada bridge ARB frontera), se requiere:

-MAC ARB destino (u origen), puertos pertenecientes a la instancia de árbol y roles de los mismos (rol de puerto raíz o de puerto designado).

En cada puerto se maneja un vector de prioridad de puerto para cada instancia de árbol. Este vector contiene los siguientes campos:

- *bridge_id_raíz: ID del Bridge raíz de la instancia*

- *coste_a raíz:* coste del camino al bridge raíz de la instancia de árbol

- *bridge_id_designado:* identidad del bridge designado.

- *id_puerto_designado*: identidad del puerto designado (emisor de la BPDU).

- *id_puerto_receptor*: identidad del puerto designado (receptor de la BPDU).

Las instancias de árbol resultan en una tabla de encaminamiento de la forma: numero de instancia, bridge_id del bridge origen del árbol (raíz), ID de puerto raíz, ID de puerto designado 1....puerto designado n. Siendo k el numero medio de puertos por bridge, el número de instancias es igual al numero de bridges raíz que coincide con el número de bridges frontera, el cual es menor o igual que M. La tabla de encaminamiento de todas las instancias es proporcional a M * k. El consumo de memoria para los árboles es pues O (M) en cada ARB.

6.2.3 Complejidad de mensajes AMSTP.

La complejidad de los mensajes de AMSTP es similar a la de protocolos como MSTP considerando un número de instancias de árbol igual al número de bridges frontera. Dado que las BPDUs de todos los árboles se transmiten agrupadas en una única BPDU, el número de mensajes es igual que el de RSTP, pero con tamaños superiores por incluir las distancias de todos los árboles del troncal en la BPDU compuesta. La longitud de mensaje es *O(BF)* (siendo BF el número de bridges frontera). La frecuencia de los mensajes es la misma que en RSTP y coincide con el *hello time* de los bridges. El procesado de cada BPDU agrupada en el ARB actualiza todas las instancias de árbol de dicho ARB.

6.2.4 Caminos mínimos AMSTP en troncal

Los caminos de la red troncal con el protocolo AMSTP siguen siempre las ramas de un árbol de expansión cuya raíz es el bridge frontera origen de la trama, el bridge por donde la trama ingresa en el troncal, terminando la trama su recorrido en el troncal en el bridge frontera de salida. Estos caminos del troncal construidos por AMSTP son mínimos.

Demostración. El protocolo de formación de instancias de árboles de AMSTP opera de forma tal que en cada bridge se compara el coste hasta el bridge raíz y se elige como puerto raíz hacia el mismo aquel cuyo coste al raíz es menor:

coste_a_bridge_raíz via puerto j < coste_a_bridge raíz via puerto 1,.....,P (excepto j)

Si existe un camino de menor coste a bridge raíz desde el bridge frontera destino por otro puerto, este puerto será elegido puerto raíz, por lo que el camino es siempre mínimo del raíz (bridge frontera) hasta el destino. Esto se cumple en todos los bridges del camino hacia arriba hasta el bridge frontera origen, por lo que el camino es óptimo.

6.2.5 Topologías aleatorias de red troncal

El protocolo AMSTP utilizado en el troncal es de camino mínimo, por lo que, aparte de compararlo con RSTP, como se hace al evaluar la capa de acceso, para valorar su comportamiento en redes troncales se trata de calcular los caminos mínimos en las redes troncales objetivo. Con vistas a evaluar el rendimiento del protocolo AMSTP surge la cuestión de las características topológicas a suponer en las redes troncales. En primer lugar si deben suponerse aleatorias o no. Aunque poderosas razones hacen que el diseño de troncales no presente topologías aleatorias, se han simulado a fin de obtener los valores medios esperables, de forma complementaría a la evaluación de topologías regulares. Las topologías regulares permiten una valoración más sistemática del rendimiento esperable de una red según sus características y grado de conectividad de los nodos. Las topologías aleatorias no parecen recomendables para troncales por la falta de control sobre las prestaciones del troncal, entre ellas longitud máxima de camino, aspecto que debe tenerse en cuenta por razones de configuración de los árboles de expansión, aunque no es limitante.

Mediante la herramienta gt-itm, se han generado topologías aleatorias no jerárquicas de 9, 16, 25, 36 y 64 nodos con distintos grados medios de conectividad de nodos. Denominamos *grado medio* de la red al número medio de enlaces por nodo Se han generado para cada topología 10 topologías distintas por grado para grados medios de 3, 5 y 7. Mediante una rutina de la herramienta se han computado los caminos mínimos resultantes entre nodos, obteniéndose las longitudes media de caminos de cada grupo de topologías según número de nodos. Las rutas se han calculado primeramente usando métrica de número de saltos y calculando la longitud del camino en número de saltos. El grado medio de las redes se ha variado para ver la influencia en las longitudes de caminos, significativo en el diseño del núcleo de redes campus. El intervalo de confianza para un nivel de 95% es de tamaño máximo para las topologías de grado medio 3 de nueve nodos, con un valor de +/- 0,93 saltos. El intervalo de confianza se reduce

considerablemente al aumentar el número de nodos y el grado, llegando a +/- 0,47 saltos para 64 nodos y grado medio 7. En la Figura 60 se muestran los resultados. Esto permite estimar, para un tamaño dado de red, el número medio de saltos conociendo el grado medio de conectividad de los nodos de la red.

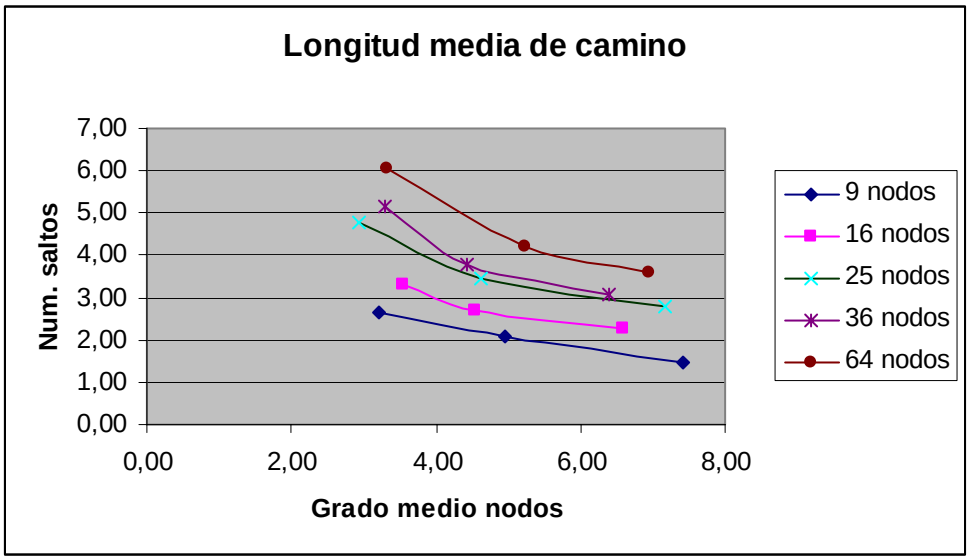


Figura 60. Longitud media de camino según grado medio de nodos.

Mediante el adecuado dimensionado del grado medio de los nodos puede controlarse el número medio de saltos. En cuanto al número máximo de saltos o longitud máxima de camino, por tratarse de topologías aleatorias esto no puede controlarse de forma eficaz. Si se desea un control total del camino máximo en el troncal deben utilizarse topologías regulares de alta conectividad, lo cual no es problema en las redes campus, y sí lo es en redes metropolitanas, por el alto coste de los enlaces entre los nodos del troncal.

En la Figura 61 se muestran las longitudes máximas promedio de camino en función del grado medio de los nodos de las topologías. Para topologías de 36 y 64 nodos el número de saltos se eleva considerablemente cuando el grado medio es inferior a cuatro.

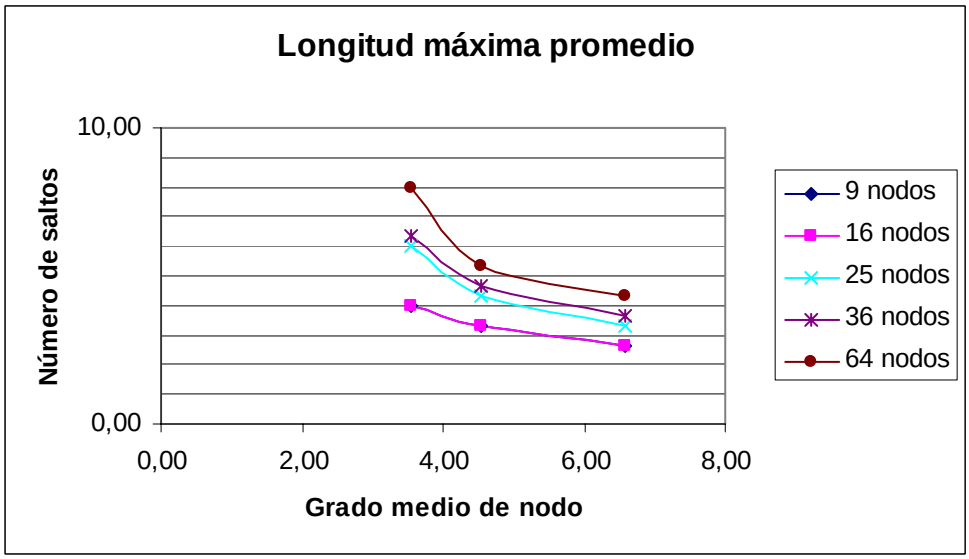


Figura 61. Longitud máxima promedio de camino según grado medio

Para una determinada longitud media de camino deseada en el troncal, el grado medio de los nodos debe aumentarse a medida que la red crece. Esto es paralelo a lo que se muestra para las topologías regulares (hipercubos).

En la Figura 62 se muestra la longitud media normalizada de caminos (computada por la herramienta como *ll media*, calculando las rutas mediante métricas de distancia euclídea entre nodos en longitud y expresando el coste del camino también en distancia).

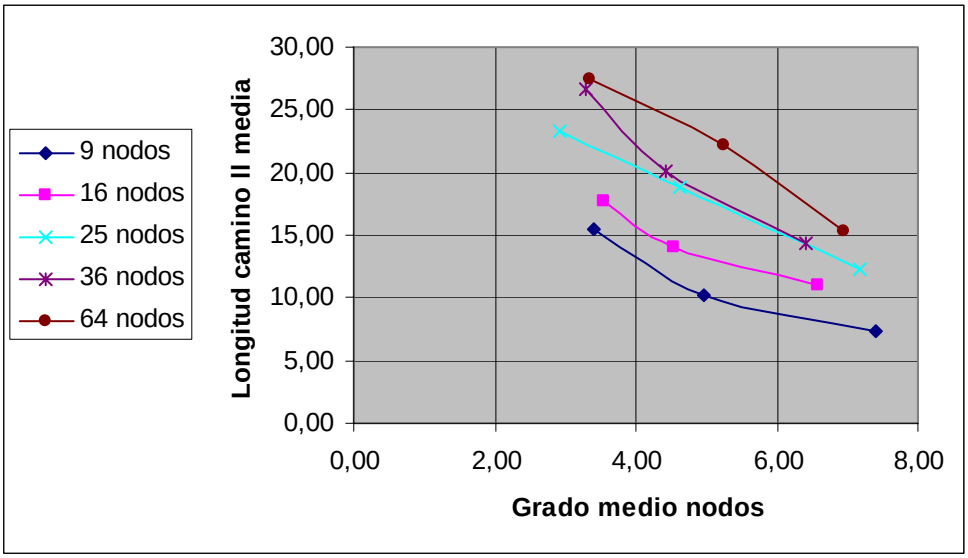


Figura 62. Longitud media de caminos respecto a grado medio.

A continuación comparamos las prestaciones de AMSTP con las implementaciones alternativas en cuanto a longitudes de camino en troncal. Por ser de camino mínimo, AMSTP presenta idénticas prestaciones (óptimas) que cualquier protocolo de camino mínimo utilizando la métrica que se prefiera en los enlaces. La comparación de longitudes de camino de AMSTP con el protocolo de Árbol de Expansión estándar actualmente utilizado, se muestra en la sección de Nivel de Acceso, donde se comparan las longitudes medias de caminos con árbol de expansión respecto a nodos con conectividad total por caminos mínimos.

6.2.5.1 Tráfico de saturación

Con el fin de obtener las capacidades máximas de cursar tráfico de las redes troncales de este tipo se han generado mediante la herramienta BRITE [BRITE] topologías aleatorias de 9, 16, 25 y 36 nodos de grado medio 4, con 17, 31, 24 y 35 enlaces respectivamente, generando con los valores por defecto de BRITE cinco topologías aleatorias de cada tipo. Los nodos se interconectan mediante enlaces de 10 Gbps, inyectando tráfico uniforme en cada nodo, incrementando dicho tráfico hasta producir la saturación de la red. Se suponen enlaces de 1 km. de longitud y tiempos de propagación en la fibra de 5 μs/km. Todos los conmutadores son 10 GE *wirespeed.* Suponemos tráfico uniforme, en el que en cada conmutador inyecta tráfico en el troncal a tasa constante. Se ha simulado dichas topologías mediante *ns2* con tráfico de entrada en cada nodo de 1, 2, 4, 8 y 16 Gbps. El simulador no permite simular a 10 Gbps, por lo que todos los parámetros de red se escalan con un factor de 1000 y se simula a 10 Mbps. No se aplica ningún control de congestión en los nodos dado que lo que se desea obtener es

la capacidad de transporte máxima de las redes. En la Figura 63 se muestran los valores medios del tráfico cursado para en función del número de nodos. Para un nivel de confianza del 95% el intervalo de confianza es menor de 10,22 % expresando el tráfico cursado como porcentaje de tráfico respecto al tráfico inyectado. Exceptuando la topología de 9 nodos, a partir de 4 Gbps inyectados por cada nodo se producen pérdidas significativas de tráfico, las redes saturan aproximadamente al inyectar 8 Gbps por ARB y el tráfico cursado máximo por enlace es menor de 7 Gbps.

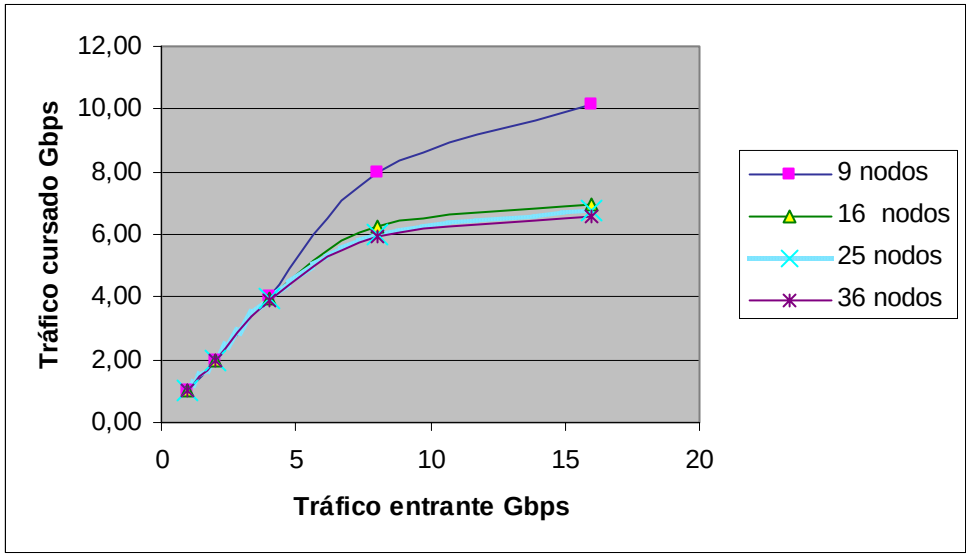


Figura 63. Tráfico cursado en topologías aleatorias

6.2.6 Topologías regulares

Se han evaluado y simulado topologías regulares de media y alta conectividad. Las topologías son de malla abierta para conectividad media, más adecuada en entornos de redes metropolitanas y de tipo hipercubo para conectividades altas. En las primeras se simularon para comparar el rendimiento del protocolo AMSTP de troncal con otros protocolos como RSTP en topologías planas regulares. No se mencionan las topologías en anillo por su inadecuación a redes campus. Las topologías en anillo se contemplan en redes metropolitanas bajo determinadas circunstancias, pero en redes campus presentan restricciones inaceptables y se congestionan mucho antes.

6.2.6.1 Malla abierta

Se ha comparado el rendimiento de AMSTP en la red troncal con el rendimiento del protocolo de árbol de expansión RSTP sobre topologías de malla abierta de los tamaños siguientes : 9, 16, 25 y 36 nodos. Se comparan retardo y máximo tráfico de salida, de forma similar l realizado en [Garcia03], pero aplicado al entorno de redes campus en cuanto a número de nodos y longitudes de enlaces entre nodos. Las topologías comparadas se muestran en la Figura 64. Se suponen enlaces de 1 km. de longitud y tiempos de propagación en la fibra de 5 μs/km. Se envían tramas de 1000 bytes a 1 trama por microsegundo, lo que equivale a un tráfico inyectado en cada nodo de 8 Gbps. Todos los conmutadores son 10 GE *wirespeed.* Suponemos tráfico uniforme, en el que en cada conmutador inyecta tráfico en el troncal a tasa constante. No se modela control de

congestión dado que se trata de obtener una comparación de rendimiento entre topologías.

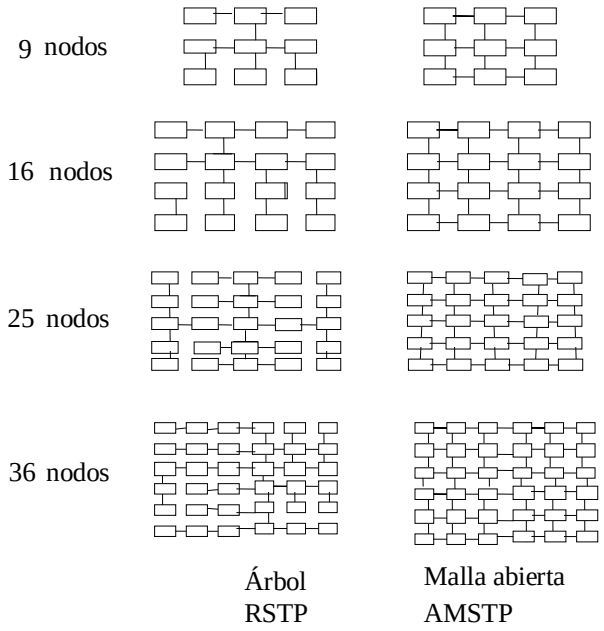


Figura 64. Topologías de red comparadas

Retardo medio en el troncal.

El retardo medio en el troncal se muestra en la Figura 65. El principal factor que contribuye al retardo es el tiempo de propagación en los enlaces ópticos, de unos 5 microsegundos por enlace. La mejora del retardo medio es consecuencia de la menor longitud de caminos de AMSTP.

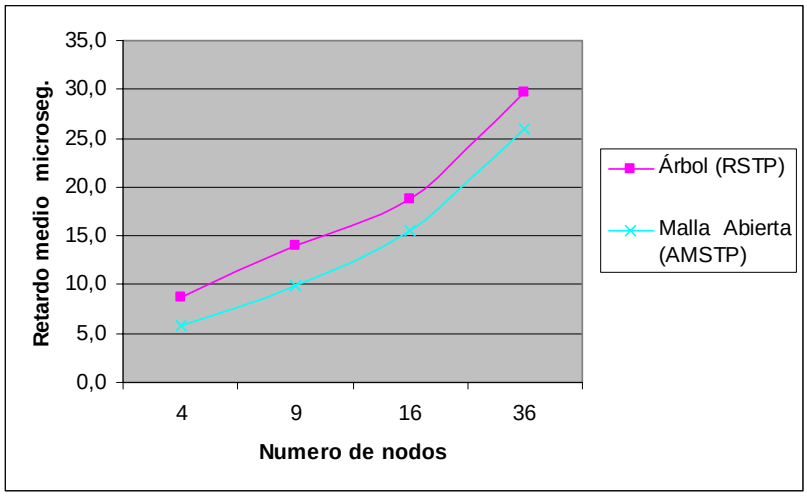


Figura 65. Retardo medio en troncal

Tráficos de saturación RSTP y AMSTP

Se ha comparado también el tráfico de saturación (tráfico máximo ofrecido) de RSTP frente a AMSTP sobre redes malladas abiertas regulares como las de la Figura 64. En la Figura 66 se muestra la saturación de tráfico con AMSTP para las topologías de 9, 16, 25 y 36 nodos. En la Figura 67 se muestra el tráfico de saturación con RSTP para las mismas topologías. AMSTP con múltiples árboles de expansión sobre topología de malla abierta se satura más tarde y escala mejor que el árbol único de expansión RSTP debido a la mejor utilización de la red. El aspecto más destacable es que AMSTP escala mejor que RSTP a partir de 16 nodos debido a la mayor distribución del tráfico y mejor utilización de la red. RSTP no es eficiente en troncales de más de 16 nodos. Pero AMSTP también sufre saturación en las topologías mayores de malla abierta regular debido al bajo grado medio de conectividad.

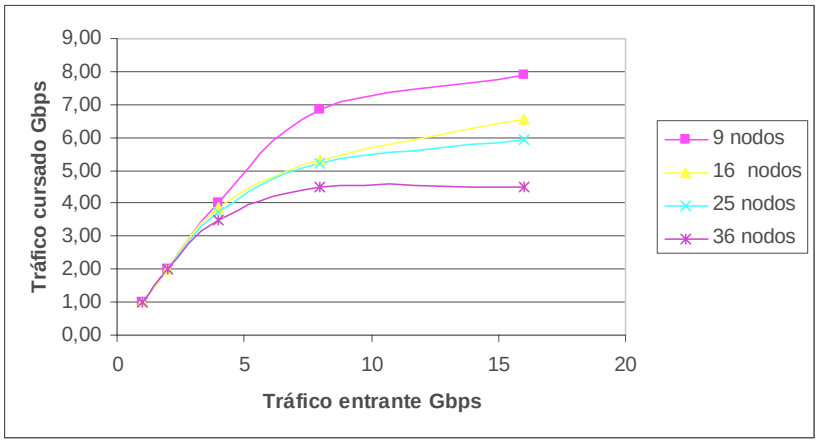


Figura 66. Tráfico cursado respecto a tráfico de entrada en troncal

Tabla 5. Tráfico cursado respecto a tráfico de entrada en troncal

Gbps			Nodos	
Tráfico cursado	9	16	25	36
1	1,00	1,00	1,00	1,00
2	2,00	2,00	2,00	2,00
4	4,00	3,88	3,75	3,49
8	6,84	5,31	5,20	4,50
16	7,91	6,56	5,92	4,48

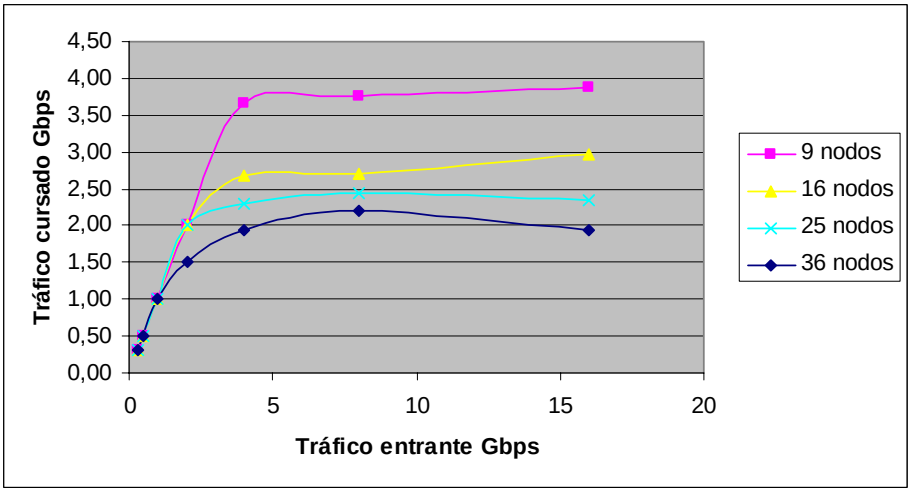


Figura 67. Tráficos de saturación con RSTP.

6.2.6.2 Topologías regulares de alta conectividad.

En los análisis de rendimiento anteriores de topologías regulares se ha supuesto conectividad en dos dimensiones. Entre ellas, las topologías de malla abierta son las más eficientes económicamente para redes metropolitanas debido al alto coste que supone una elevada conectividad mediante interconexiones de fibra óptica entre nodos distantes. En el caso que nos ocupa de redes campus el coste de la conectividad adicional es mucho menor, dado que las longitudes de los enlaces de fibra son del orden de kilómetros como máximo y dentro de un recinto privado en vez de decenas de kilómetros en un espacio no privado. Por ello en redes campus se pueden considerar topologías regulares de alta conectividad. Hemos evaluado el rendimiento de unas cuantas topologías n-dimensionales de 8, 16 y 32 nodos ($N=2^n$).

La topología elegida se denomina *red n-cubo k-ario*. Una red *n-cubo k-ario* se define [DYN97] como una red de interconexión con *n* dimensiones y *k nodos* en cada dimensión. Cada nodo en dicha red se identifica por su posición en cada dimensión, representada por el vector ($x_1, x_2, x_3, \ldots, x_n$). Dos nodos ($x_1, x_2, x_3, \ldots, x_n$) y ($y_1, y_2, y_3, \ldots, y_n$) son vecinos entre sí, si y sólo si existe un *i* tal que $x_i = (y_i + 1) \bmod k$ y $x_j = y_j$ para todo $i \neq j$. Cuando $k = 2$, cada nodo tiene *n* vecinos, que es la topología elegida, en la que el grado de todos los nodos es *n*. (Si $k > 2$, cada nodo tiene $2n$ vecinos).

Las topologías son las siguientes: cubo de 8 nodos, hipercubo y "doble" hipercubo para 16 y 32 nodos como se muestra en la Figura 68

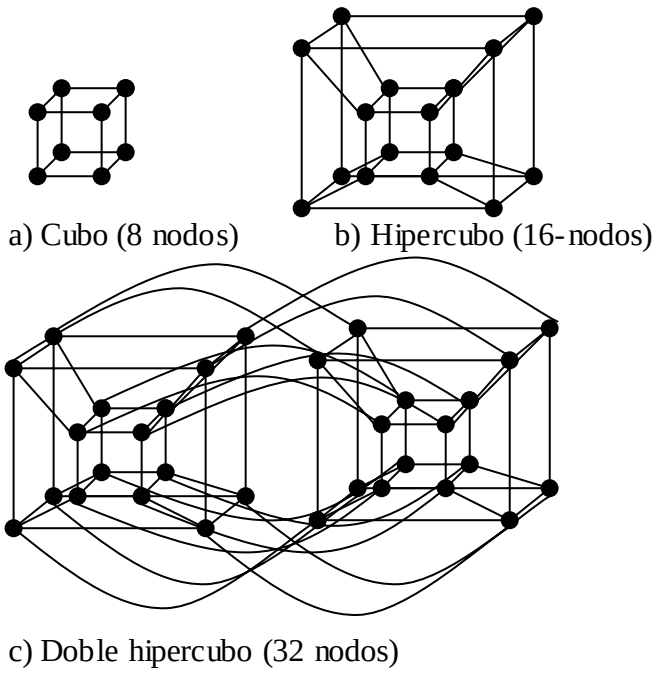


Figura 68. Topologías de alta conectividad (hipercubo).

Al contrario de lo que sucede en las topologías anteriores con RSTP y otros protocolos, y como se muestra en la Tabla 6, el tráfico ofrecido máximo (saturación) con

AMSTP aumenta con el número de nodos en vez de decaer fuertemente [García03]. Los resultados se muestran relativos a un tráfico de 8 Gbps correspondiente a λ=1.

Tabla 6. Rendimiento AMSTP en topologías de alta conectividad

	Long. Media de Camino	Longitud Máxima de Camino	Tráfico ofrecido máximo (λ)
Cubo 8 nodos	1,71	3	2,19
Hipercubo 16 nodos	2,30	4	2,34
32 nodos	2,58	5	2,42

Esto se debe al aumento de enlaces por nodo, como se detalla en los epígrafes siguientes. Por ello en la práctica el grado medio del troncal puede ser un parámetro de diseño utilizable para el adecuado dimensionado de la red. Dentro del diseño de núcleos de red de este tipo de topologías, el grado permite configurar el tráfico de saturación máximo.

No se han obtenido datos de saturación con RSTP para las topologías de alta conectividad, pero los porcentajes de utilización de la red comparados con AMSTP, como se muestra en el epígrafe siguiente, son peores que para las topologías malladas regulares debido a su mayor conectividad desaprovechada por RSTP. Por otra parte, el árbol de expansión de una malla de 16 nodos no se diferencia drásticamente del de un hipercubo. El nodo raíz es de grado cuatro en ambos casos. Por todo ello, es de esperar que las características de saturación de RSTP sobre topologías hipercúbicas no varíen mucho respecto a la topología de malla, dado que el factor limitante serán los enlaces cercanos al bridge raíz del árbol, que consolidan el tráfico de los nodos inferiores a ellos en el árbol.

6.2.6.3 Características de las topologías regulares

Para completar la comparación de topologías regulares, veamos el rendimiento esperable de otras dos topologías regulares: las topologías en anillo y en estrella. En la Tabla 7 se muestran las características relativas a número de enlaces y número de saltos de las topologías más simples en dos dimensiones: anillo, estrella y malla abierta.

Tabla 7. Propiedades de topologías simples

Topología	***Enlaces***	***Diametro (saltos)***
Anillo	***n-1***	***n/2***
Estrella	***n-1***	***2***
Malla	$2*(n-\sqrt{n})$	$2*(\sqrt{n}-1)$

La topología en estrella presenta la característica de diámetro mínimo de red usando el mínimo número de enlaces (n-1) capaz de mantener conectada la red. Aunque el fallo de un enlace solamente afecta a un nodo, el fallo del nodo central provoca el colapso de

la red. La duplicación de este nodo central supone duplicar el número de enlaces, alterando todas sus características, pero mejora el problema de fiabilidad. La topología en anillo presenta en cambio la máxima disponibilidad, igual en todos los nodos, para el mínimo número de enlaces (n-1) que mantiene la red unida. Esta característica de la topología en anillo es una de las razones de su empleo en el pasado en transmisión óptica (SDH) y redes locales (FDDI, Token Ring). Pero el diámetro de la red es elevado y aumenta con el número de nodos, lo que indirectamente implica que la red se sature antes al recorrer el tráfico mayor número de tramos en el núcleo antes de abandonarlo y sumarse por tanto mayor número de flujos en cada enlace. Se han excluido por esta razón de las comparaciones de este capítulo las topologías y protocolos que operan sobre anillos, como Resilient Packet Ring, por no adecuarse a las redes campus. El número de enlaces y el diámetro resultantes de las expresiones de la Tabla 7 para los tamaños de red considerados se muestran en la Tabla 8.

Tabla 8. Características de topologías simples.

Numero de enlaces

Num. Nodos	Anillo	Malla abierta
4	3	4
9	8	12
16	15	24
25	24	40
36	35	60
64	63	112

Diametro de la red

Num. Nodos	Anillo	Malla abierta
4	3	2
9	8	4
16	15	6
25	24	8
36	35	10
64	63	14

Como puede verse, el diámetro de la red crece rápidamente para el anillo y para la malla abierta a partir de 25 nodos. Esto es consecuencia de la baja conectividad y constituye el lado aspecto negativo del ahorro de infraestructura usando pocos enlaces.

Para las topologías de alta conectividad hipercúbicas, el número de enlaces y diámetro de las mismas son respectivamente las que se muestran en Tabla 9 :

Tabla 9. Características de topologías hipercúbicas ($n = 2^m$).

Num. Nodos	Num. Enlaces	Diametro red	grado m
8	12	3	3
16	32	4	4
32	80	5	5
64	192	6	6

Como puede verse, en estas topologías el diámetro de la red coincide con el grado de la red y crece de la misma forma lineal. Tiene valores menores de 7 para las topologías consideradas (hasta 64 nodos). Esta topología visualiza claramente los compromisos necesarios en el diseño de troncales de redes campus cuando es posible utilizar toda la infraestructura como es el caso de AMSTP. El número de enlaces necesarios aumenta

rápidamente con el número de nodos, lo que no es problema en nuestro caso de redes campus y sí lo es en el caso de redes metropolitanas.

La aplicación de estos valores a la comparación de la utilización de la red troncal por AMSTP y RSTP en las topologías de malla abierta y topologías hipercúbicas da los siguientes resultados: la utilización de la red troncal con AMSTP es del 100% en todos los casos en ambas. Con RSTP la utilización de la red varía para las topologías de malla abierta entre el 67% para 9 nodos y el 56 % para 64 nodos y para las topologías hipercúbicas entre el 58 % para el cubo de 8 nodos y el 33 % para 64 nodos. La menor utilización de la red se debe al mayor grado de conectividad de las topologías hipercúbicas.

6.2.7 Conclusiones red troncal

Se puede concluir con respecto al rendimiento de AMSTP, que en cuanto a longitud de caminos medio y máximo se comporta de forma óptima y con mucha menor sobrecarga y complejidad de proceso que un protocolo de estado de enlaces por ser un protocolo de árbol de expansión, equivalente a la de N árboles de expansión. Frente a RSTP utiliza la red de forma mucho más eficiente sin dejar enlaces inutilizados y ofrece características de saturación muy superiores que permiten aumentar la capacidad del troncal aumentando el grado de conectividad de los nodos de la red. Es aplicable a todo tipo de topologías, a diferencia de RSTP, con topología activa de árbol u otros, como RPR, orientados a anillos. En cuanto a escalabilidad, su bajo coste computacional y el reducido ancho de banda consumido por las BPDUs respecto a las capacidades del troncal lo hacen aplicable al núcleo de redes campus y suficientemente escalable para aplicarse en redes de muy alta capacidad formadas mediante conmutadores de media capacidad fuertemente interconectados.

6.3 Nivel de Acceso.

Se analiza a continuación el rendimiento en la red inferior que actúa como capa de Acceso, en la que los protocolos ejecutados son: RSTP en el escenario básico, RSTAR ó RSJ cuando se optimizan caminos en el nivel inferior.

Los caminos en la red inferior se dividen en dos tipos: caminos intra-área y caminos inter-área. Los caminos intra-área se establecen mediante el protocolo de Árbol de Expansión mejorado con el protocolo STAR o alternativamente con el protocolo RSJ (protocolo jerárquico RSTASTAR). Los caminos inter-área de la red inferior son siempre entre el bridge raíz ARB y otro bridge de su árbol de expansión. En los árboles de la capa inferior los caminos nacen (descendentes) o mueren (ascendentes) en el bridge raíz, siendo esta distancia minimizada por el protocolo RSTP.

6.3.1 Caminos intra-área en STAR y RSJ

En [Lui02] se demuestra que el protocolo STAR siempre mejora o mantiene el camino de árbol de expansión, reemplazándolo por un camino alternativo al árbol de expansión de menor coste cuando es posible, porque su algoritmo asegura dicha mejora aun a costa de ignorar mejoras que estadísticamente podrían serlo, pero a riesgo de empeorar en ocasiones el coste del camino resultante. El protocolo STAR conoce a veces solamente los costes máximos y mínimos de camino porque opera con estimaciones. Estas estimaciones se deben a que existen bridges 802.1D en la red que no ejecutan STAR y no siempre es posible obtener la distancia exacta entre bridges STAR. En el teorema 1 se establece que:

Dados dos bridges STAR k y x, al finalizar el proceso de búsqueda de caminos, si el camino en la tabla de distancias DVT(k,x) no es un camino de árbol, entonces la distancia(k,x) < distancia_árbol(k,x).

Mediante varios lemas auxiliares relativos a la obtención del camino alternativo, se demuestra el teorema antedicho. Dichos lemas expresan básicamente que en la tabla de distancias un camino nunca es reemplazado por uno de coste superior mediante cualquiera de los mecanismos de descubrimiento.

Los caminos establecidos con el protocolo RSJ también mejoran siempre el camino respecto al protocolo de Árbol de Expansión. La primera razon es que RSJ se basa en el protocolo STAR y opera con los mismos principios de comparación de costes respecto al coste del camino por el árbol de expansión. No se elige nunca un camino que tenga el riesgo de ser más largo que el reemplazado.

6.3.2 Protocolo RSTP

El protocolo RSTP está descrito en el estándar IEEE 802.1D [802.1D04] por lo que no nos detendremos en el análisis de su rendimiento. Para dar una idea de la sobrecarga del protocolo, hay que considerar que básicamente cada nodo envía BPDUs cada *Hello time* a sus nodos vecinos. El número de mensajes total es pues *O(E)* siendo *E* el número de enlaces. El ancho de banda consumido por las BPDUs (34 octetos transmitidos cada dos segundos), es muy reducido comparado con la capacidad de los enlaces Ethernet.

6.3.3 Protocolos STAR y RSTAR

Se analizan a continuación las prestaciones y recursos requeridos por STAR y RSTAR. El rendimiento del protocolo STAR está descrito en [Lui02] y aquí solamente lo resumimos dado que RSTAR es básicamente una migración de STAR.

6.3.3.1 Complejidad de mensajes

En el protocolo STAR, los mensajes adicionales a IEEE 802.1D tienen las funciones siguientes: el mantenimiento de los enlaces cruzados, la obtención de los vectores distancia y la identificación de los bridges agentes de los sistemas finales.

-Mantenimiento de los enlaces cruzados: En STAR se usa protocolo *Hello* en los enlaces cruzados para detectar el fallo del enlace cruzado. Como estos mensajes son punto a punto y no se propagan más allá del bridge vecino, la sobrecarga respecto al protocolo de árbol de expansión es nula.

-En cuanto al cálculo de los vectores distancia, el coste máximo está acotado por el diámetro de la red multiplicado por el número de mensajes de establecimiento del árbol de expansión. Una vez elaboradas las tablas de encaminamiento, no se transmiten más mensajes STAR, excepto las BPDUs periódicas de mantenimiento del árbol de expansión, por lo que el número de mensajes en una red estable es despreciable. Si hay un fallo de un enlace cruzado hay que recalcular los vectores distancia, por lo que se envía una notificación por todo el árbol para disparar el proceso de recálculo. Por lo tanto el coste de mantener los enlaces cruzados es el mismo que el requerido por las BPDUs de cambio de topología del protocolo RSTP.

6.3.3.2 Longitud de caminos

Los caminos establecidos por el procotolo RSTAR son mejores que los de árbol de expansión pero no son óptimos. Esto es debido a que normalmente solamente una fracción de todos los bridges de la red será capaz de ejecutar RSTAR y podrá por tanto establecer caminos alternativos. Pero dichos caminos son siempre iguales o mejores que los caminos vía árbol de expansión, siendo la mejora algo mayor cuanto menor es el grado medio de los nodos de la red (porque un enlace cruzado supone mejor acortamiento de camino relativo cuanto menor conectividad hay en la red y más largos son los caminos por el árbol de expansión) y también cuanto mayor es el porcentaje de bridges RSTAR en la red (es decir, existe un mayor número de enlaces cruzados alternativos al árbol de expansión). Como dato indicativo, el valor máximo de la relación entre la longitud de camino vía árbol y vía enlace cruzado es de alrededor de 5 [Lui02].

Es importante destacar que esta no optimalidad de caminos de RSTAR mencionada en el párrafo anterior es aplicable solamente al tráfico local dentro de la región o capa de Acceso (tráfico intraárbol). El tráfico inter-área que entra o sale del troncal sigue siempre un camino óptimo porque utiliza el árbol de distribución de la red inferior desde o hacia su bridge raíz (ARB). Esto hace que los caminos inter-área sean óptimos de extremo a extremo dado que tanto los tramos sistema final-ARB como ARB-ARB lo son. La Figura 73 ilustra esta característica.

6.3.3.3 Utilización de infraestructura

RSTAR incrementa la utilización de la infraestructura en proporción al número de enlaces cruzados establecido entre bridges RSTAR. Este número depende de forma directamente proporcional al número de enlaces entre bridges STAR que presenten costes ventajosos respecto al Árbol de Expansión, respecto al total de enlaces de la red. Este número de enlaces entre bridges STAR depende de la estrategia de colocación de los bridges STAR. Si se eligen mal la ventaja aportada por los bridges STAR puede ser mínima o nula. Intuitivamente puede deducirse que los bridges STAR deberán estar unidos a otro u otros bridges STAR de forma “horizontal”, en lugar de “vertical” (por el árbol de expansión). Para una utilización ventajosa de la infraestructura es pues necesaria una colocación efectiva de los bridges RSTAR y que los costes de sus enlaces sean ventajosos frente a los del árbol de expansión. La estrategia de posicionamiento de bridges STAR descrita en [Lui02] es aplicable aquí. Dada una topología de red y un número de bridges STAR, se plantea como un problema de optimización la elección del

conjunto de bridges a reemplazar por STAR que genere el mayor número de enlaces cruzados ventajosos.

6.3.4 Protocolo RSJ

El protocolo RSJ deriva del protocolo RSTAR complementado a su vez con encaminamiento jerárquico basado en las coordenadas RSTAA. Incorpora las ventajas del encaminamiento jerárquico sin los inconvenientes de la complejidad de configuración asociada a los protocolos jerárquicos. Dado que RSJ encaminamiento por defecto por el árbol de expansión, la problemática habitual de los protocolos de vector distancia queda muy atenuada. Frente a los protocolos de estado de enlaces, que normalmente requieren información completa de la topología, los de vector distancia pueden trabajar con información simplificada (abstracción de la topología) conservando la simplicidad.

6.3.4.1 Almacenamiento

Siendo *B* el número de Bridges RSJ, *SF* el número de sistemas finales, las necesidades de almacenamiento de RSJ en los bridges AB son las siguientes:

-Tabla de encaminamiento (*TE*) en cada bridge AB. Esta tabla tiene una entrada por cada bridge, no por cada sistema final, dado que a diferencia de STAR, RSJ no encamina por sistemas finales sino por bridge AB. *TE = O(B).*

-*TSF*. Tabla de sistemas finales del AB: *ID puerto, MAC de hosts conectados al puerto*. Su tamaño es el máximo de sistemas finales por AB: *O(SF/B)*

-*TVD*. Tabla de vectores distancia entre bridges AB. Contiene: *MAC bridge destino, Coordenada RSTAA de bridge destino, Distancia Mínima a bridge Destino, Dist. Máxima., puerto de salida para destino, MAC del bridge AB (siguiente salto) y Relación entre ABs (antecesor, sucesor, ninguna relación), profundidad (profundidad relativa si hay relación).*

6.3.4.2 Complejidad de Mensajes

Se analiza a continuación la complejidad de mensajes. El número de mensajes por nodo en el caso de vector distancia es *O(d)* siendo *d* el grado medio, mientras que en los protocolos de estado de enlaces es *O(n*E)* siendo *E* el número de enlaces dado que se reciben mensajes de todos los nodos.

RSJ es básicamente un protocolo de vector distancia que complementa al de Árbol de Expansión (RSTP). En el protocolo RSJ, salvo lo indicado explícitamente, la frecuencia y contenidos de mensajes son equivalentes a los de RSTAR, por lo que se evalúa su complejidad relativa a la de STAR y RSTAR estudiando las diferencias con dichos protocolos:

-RSTAR y STAR no recalculan vectores distancia hasta que se produce un cambio de topología [Lui02]. Esto reduce a cero sus mensajes en situación de estabilidad, exceptuando los del protocolo de Árbol de Expansión. RSJ tiene la opción de actualización permanente e inmediata al cambiar un nodo su vector distancia, pero la difusión está limitada por el algoritmo (radio de difusión). Si dicha opción está activada,

se intercambian vectores distancia entre bridges RSJ periódicamente o al producirse cambios de topología. En este caso el número de mensajes es menor o igual que el de un protocolo de vector distancia.

-El tamaño de los vectores distancia se reduce por el alcance limitado de la difusión de las rutas. Los valores de radio de difusión para redes de grados medios entre 3 y 4 a aplicar son entre 2 y 4. Las redes de grado dos son casi árboles de expansión, por lo que apenas deshabilitará enlaces RSTP y no procede el uso de RSJ ni RSTAR. Se considera efectiva una reducción de rutas difundidas hasta el rango del 30%-70%. Por el encaminamiento jerárquico la tabla es de tamaño *O(log(n))*, similar a Landmark [Tsu88].

-Localización de sistemas finales. El servidor ARP resuelve a la vez el AB destino y la dirección MAC del sistema final destino. Esto supone un coste de dos mensajes adicionales unicast (petición y respuesta del AB al servidor) y se evita la emisión de un mensaje en broadcast. Por tanto la localización de sistemas finales (resolución del AB destino) supone para el bridge un mensaje adicional al servidor y el mensaje de contestación al ARP en unicast al sistema final origen.

Establecimiento de coordenadas RSTAA

El establecimiento de las coordenadas RSTAA se produce conjuntamente con el establecimiento y reestablecimiento del árbol de expansión RSTP. Salvo en el caso de cambios de topología, requiere una pequeña sobrecarga adicional sobre las BPDUS periódicas de RSTP y de su misma frecuencia. Se transmite una BPDU RSTAA cada *Hello time.*

6.3.5 Longitudes de caminos intra área.

Se analizan en este epígrafe los caminos intra área de los protocolos del nivel inferior (capa de Acceso): RSTP, RSJ y se comparan con el camino mínimo suponiendo conectividad total, ningún enlace deshabilitado) para el tráfico local (intraárea). Se han generado redes aleatorias de grado medio cuatro con la herramienta Brite [Brite] de 9, 16, 20 y 36 nodos. Se han computado y comparado sus longitudes medias de camino. Se comparan las redes usando conectividad total (caso de encaminamiento por camino mínimo), conectividad usando un 50 % de los bridges con el protocolo RSJ y conectividad con árbol de expansión (RSTP, escenario simplificado). Se ha elegido de forma aleatoria el bridge raíz de árbol de expansión en cada caso. En la Figura 69 se muestra el número medio de saltos para las diferentes topologías y protocolos.

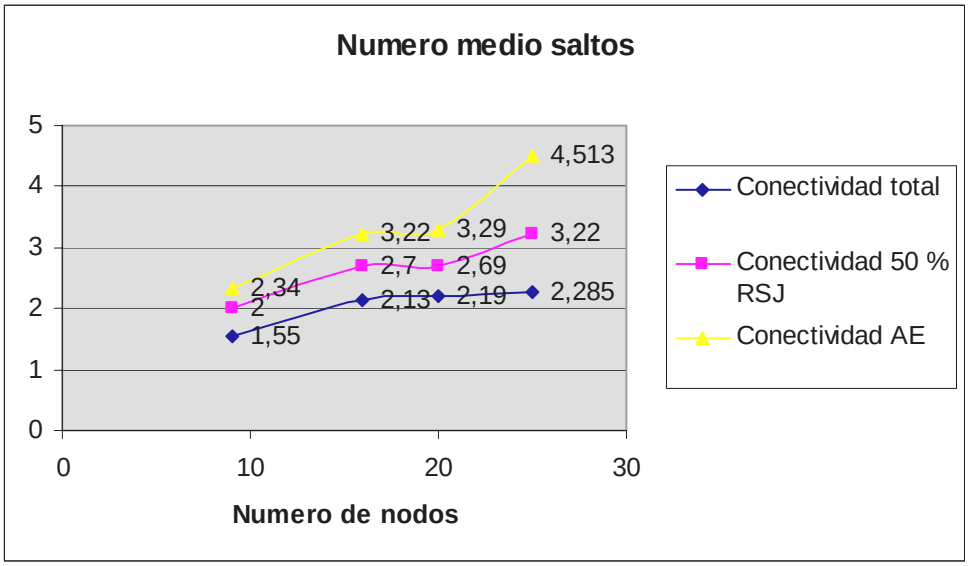


Figura 69. Longitud media de caminos intra área.

El comportamiento de la longitud máxima de camino es muy diferente al de la longitud media, como se muestra en la Figura 70. El valor máximo es independiente del protocolo en redes de pocos nodos y se dispersa al aumentar el número de nodos, de acuerdo con la conectividad mayor o menor proporcionada por el protocolo. RSJ hace descender muy ligeramente el camino máximo (aunque sí su frecuencia) respecto al árbol de expansión debido a que responde a los casos más desfavorables de encaminamiento.

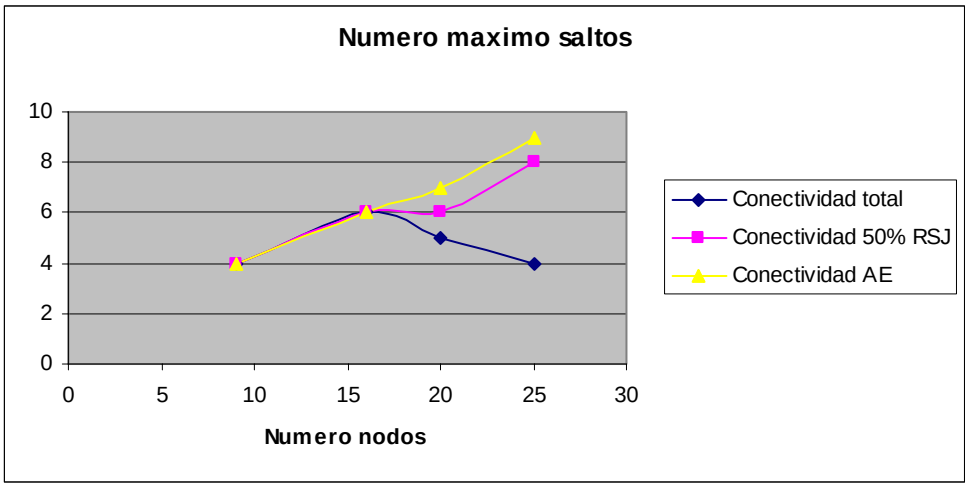


Figura 70. Longitudes máximas de camino intra área.

Por su mismo planteamiento, el rango de mejora posible de caminos obtenible con RSJ está acotado entre los caminos mínimos posibles cuando hay conectividad total (todos los enlaces activos y protocolo de encaminamiento de camino mínimo) y los caminos del árbol de expansión (AE), a los que intenta mejorar.

Tomando como referencia (1.0) la longitud media de camino de RSJ para el 50% de bridges RSJ, las longitudes medias relativas de los caminos obtenidos en el caso de conectividad total y de árbol de expansión son los siguientes: 0,78 para los caminos de conectividad total y 1,24 para los caminos de Arbol de Expansión. Al incrementar la conectividad, con índices de penetración de bridges RSJs superiores acercarán el camino medio de RSJ hacia el de conectividad total e inversamente, los inferiores lo acercarán al de árbol de expansión.

6.3.6 Eficiencia de utilización de infraestructura con árbol de expansión.

Como se ha dicho repetidamente, la utilización de la infraestructura obtenida con el protocolo de árbol de expansión estándar es baja. Esta utilización es tanto menor cuanto mayor es la conectividad e la red. En efecto: un árbol de expansión se caracteriza por tener siempre un número de enlaces igual al número de nodos menos uno:

$$e_a = n - 1$$

porque como el árbol debe unir a todos los nodos de forma única y sin bucles, cada nodo utiliza un enlace para unirse al árbol, excepto el nodo raíz.

Si consideramos el grado medio de conectividad de una red *m*, el número de enlaces total de enlaces es :

$$e = n * m / 2$$

La eficiencia de utilización de la infraestructura será:

$$E\ (\%) = 100 * e_{ae} / e = 100 * (n\text{-}1)*2 / n*m$$

Donde la relación *(n-1)/n* tiende rápidamente hacia la unidad al aumentar *n*, quedando la eficiencia inversamente proporcional a *m*. Esto es inmediato, dado que todos los enlaces que excedan el árbol de expansión son superfluos desde el punto de vista de eficiencia, no de fiabilidad.

El dilema se plantea porque si el grado de los nodos se hace bajo para obtener alta eficiencia, los caminos del árbol de expansión se alargarán, con lo que la eficiencia de utilización se consigue al precio de alargar los caminos, lejos del óptimo. Igualmente si el grado es bajo el peligro de fraccionamiento de la red en caso de fallo es mucho mayor. Es el caso de la las estructura en estrella pura, de árbol muy eficaz, pero que en caso de fallo del bridge central la estructura completa fallará. La solución consiste en equipar los enlaces redundantes adecuados que entren en servicio al fallar otros de forma que el camino medio se mantenga reducido. En la Figura 71 se muestran los grados de utilización de la infraestructura de red para diversos grados medios de la red.

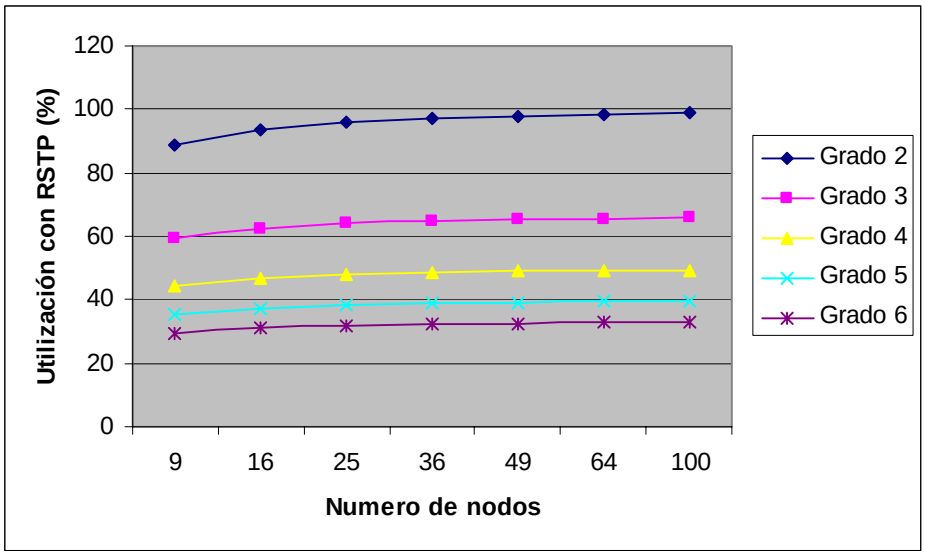


Figura 71. Utilización de la red respecto a grado medio de la red con RSTP.

El grado medio de las redes campus actuales lo estimamos entre tres y cuatro, por lo que la utilización media correspondiente de las redes campus es baja. Si se emplean topologías libres de bucles de tipo árbol, con grado cercano a dos, la utilización puede ser alta y es más económico utilizar RSTP, pero la red pierde su resistencia ante fallos de nodos o enlaces.

6.3.7 Utilización de infraestructura con RSJ

Los enlaces adicionales entre bridges RSJ mejoran la utilización de la infraestructura de red. El promedio para redes de grado medio 4 y de 9 a 25 nodos es el siguiente: el 53,4 % de los enlaces son desactivados por el protocolo de Árbol de Expansión, de los que el 34,35 % son reactivados por ser bridges entre RSJ, permaneciendo el 31.45 % desactivados.

La mejora de la utilización de la infraestructura es pues: El 31.45 de 100 de los enlaces que permanecen desactivados frente al 52 de 100 de partida supone una mejora de 52 /31,45 del 65,34 % más de utilización.

La utilización promedio de infraestructura para el 50 % de bridges RSJ es la siguiente: la utilización media de infraestructura varía desde un 100% para protocolos de camino mínimo (Djikstra) al 47,6 % para el protocolo de árbol de expansión. El protocolo RSJ obtiene en las condiciones indicadas un 68,50 % de utilización de la infraestructura, lo que supone un 43,91 % más de utilización que el Arbol de Expansión.

6.4 Tráfico en los ARB.

Se analiza a continuación la carga de tráfico a soportar por los ARB. Los ARB son el elemento clave de las redes campus propuestas. Al ser el ARB bridge raíz de la red inferior y de su instancia de árbol de la red superior, soporta todo el trafico saliente y entrante al/de los árboles locales, por lo que es el elemento a dimensionar de forma más cuidadosa. Esta aparente sobrecarga resulta en la práctica más aparente que real por dos razones: es habitual en el diseño de conmutadores la funcionalidad de agregación de tráfico en enlaces ascendentes (*uplinks)* y por otro lado, en la propuesta completa, parte del tráfico local intra área puede desviarse y no atravesar el ARB. En todos los casos, pero con mayor relevancia en la arquitectura básica, el tráfico agregado por ARB debe ser uno de los parámetros base para determinar el número de ARBs a utilizar en el núcleo.

El tráfico total no pasa por el ARB porque existe reparto de carga entre el tráfico local y de troncal (intra-área e inter-área), excepto cuando el protocolo de red inferior es RSTP. El tráfico en el ARB se reduce debido a que el protocolo RSJ desvía el trafico intra-área del Árbol de Expansión utilizando enlaces transversales más cortos. El dimensionado deberá tener en cuenta esto para optimizar la carga en la red. Por ello la distribución relativa de tráfico inter/intra-área es un elemento importante en el diseño de la red campus. También lo es un adecuado diseño de la red inferior, que con suficientes caminos alternativos al Árbol de Expansión mejorará mucho el rendimiento respecto a una red con estructura cercana a la de árbol puro, sin caminos alternativos para ser utilizados. El diseño adecuado de la red inferior permitirá desviar el tráfico intra área evitando que atraviese el bridge raíz. Esto requiere el número suficiente de bridges RSJ (o RSTAR) interconectados entre sí, pero no de forma tan directa que sean elegidos como árbol de expansión, por lo que los costes de los enlaces que los unen deberán ser menores que la suma de los costes hasta el bridge raíz de ambos y tales que no alteren el árbol de expansión del que se parte, porque replantearía la red sobre bases distintas.

El diseño más adaptado para los ARB es evidentemente uno que soporte altas capacidades en los puertos de troncal y menores capacidades por puerto en los puertos hacia el árbol, por ejemplo disponiendo de 24 puertos de 100 Mbps en el lado de distribución y de 2 a 4 puertos de 1 Gbps en el lado troncal (puertos de troncal)).

Como se indicaba al describir los ARBs en el epígrafe de Dispositivos, la funcionalidad esencial de ARB podía implementarse en la práctica de forma separada: como un ARB propiamente dicho parte del troncal y un AB conectado a él como se por

un enlace como se indica en la Figura 72. La parte de distribución funciona como un bridge RSTP a todos los efectos y la funcionalidad específica de ARB puede estar incluida en la parte de Bridge Frontera del troncal. Estrictamente hablando, si consideramos un bridge troncal frontera con un único puerto de distribución no hace falta nada más si lo unimos a un bridge RSTP unido por ese enlace, pero es necesario que el bridge ARB sea el bridge raíz de todo el árbol de la red de distribución inferior para mantener la conectividad con el núcleo, dado que el ARB actúa como enlace entre el árbol de expansión de la red inferior y la instancia de árbol de expansión superior en el troncal.

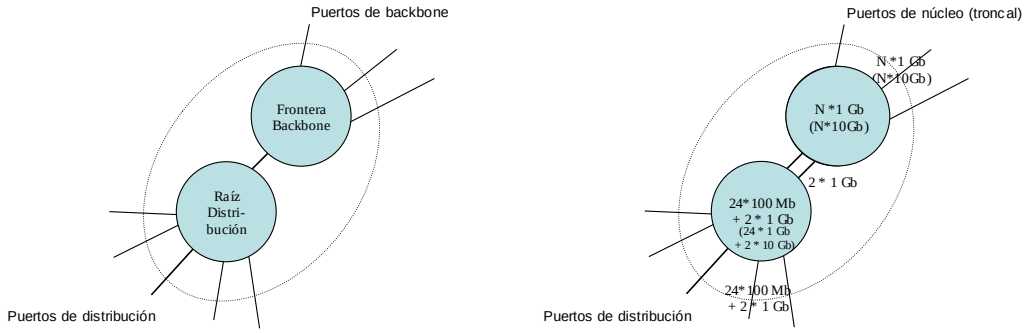


Figura 72. Implementación ARB con conmutadores N*1 y AB 24+ 2

Esta conexión es adecuada para la tipología de conmutador de 24 puertos + 2 enlaces ascendentes de agregación (*uplinks*). En la Figura 72 se muestra una implementación del ARB en formato conmutador de tipo N * 1Gbps (alternativamente N*10 Gbps) y un AB en formato 24*100 Mbps (alternativamente 24*1 Gbps+ 2*10 Gbps). Los dos enlaces ascendentes se utilizarían con agregación de enlaces para unirse al bridge frontera de troncal, el cual seria normalmente del tipo n * 1, a velocidad igual a la de los uplinks. Ocasionalmente, en troncales implementados con enlaces de capacidad modesta, podría ser del tipo 24 + 2 a velocidad inferior a la de los uplinks, pero conectado de forma opuesta. Para menor granularidad en el reparto de carga y mayor fiabilidad, puede usarse la configuración activo-activo de la Figura 37.

6.5 Servidores ARP/AB.

La utilización de servidores ARP obedece a razones de limitación de tráfico de difusión y de seguridad en la subred IP. Si por el contrario se empleara difusión de ARP convencional, el tráfico no sería excesivo en ancho de banda, pero sí en ocupación innecesaria de los sistemas finales. En [Mye04] se muestran medidas de pico y promedio de una red campus. El trafico broadcast ARP observado en un host es del orden de 0,468 mensajes ARP/seg de pico y 0,0362 ARP/seg de media por sistema final. Las medidas realizadas en el Area de Ingenieria Telemática de la Universidad Carlos III dan resultados del mismo orden: 0,04 ARPs por segundo por host de tráfico medio. Este resultado está ligado al tratamiento de ARPs en los sistemas finales [ARPWin] principalmente Windows, cuya política de caché ARP consiste en descartar las direcciones no utilizadas durante dos minutos y para las que se renuevan por estar utilizándose, retransmitir una petición ARP cada 10 minutos.

El número de accesos por segundo a los servidores ARP es función del número de sistemas finales *SF* total de la red campus. Para el cálculo de tráfico ARP, utilizaremos las medidas mencionadas [MEZ04]. El reparto de carga propuesto para los servidores ARP permite asignar un número variable de servidores según el tamaño de área y el nivel de servicio deseado. El cálculo para el tráfico a atender por los servidores ARP es: Para 20.000 PCs el tráfico de pico supone 9.364,8 mensajes ARP por segundo equivalentes a 2,4 Mbps y 362 ARP por segundo de tráfico medio equivalentes a un tráfico ARP promedio de 185 Kbps. Para los 100.000 dispositivos considerados máximos de las redes consideradas los valores se multiplicarán por cinco.

Con estas cifras es fácil ver que, si en lugar de emplear servidores se realiza difusión convencional de ARP, la sobrecarga principal no es de ancho de banda consumido en la red, relativamente bajo si consideramos que la mayoría de enlaces Ethernet será de 100 Mbps o superior, sino del tráfico a procesar por *todos* los sistemas finales de la red, tráfico inútil en los sistemas finales a excepción del sistema final al que iba dirigido el mensaje ARP, frente al resto de los sistemas finales que simplemente comprueban que el ARP recibido no va dirigido a ellos y deben descartarlo.

El empleo de servidores ARP no elimina la necesidad de mantener el mecanismo estándar de difusión ARP. Uno de los casos no cubiertos por la utilización de proxies ó de servidores ARP se produce cuando un sistema final se traslada silenciosamente de un punto de la red a otro. En este caso, si la validez de la caché no ha caducado, la respuesta del servidor o del Proxy será errónea y el sistema no será encontrado hasta que se realice un ARP con difusión.

6.5.1 Almacenamiento requerido

El almacenamiento requerido en los servidores ARP es el siguiente:

-Tabla de Sistemas Finales: por cada entrada de sistema final se guarda, como mínimo: dirección IP, dirección MAC, coordenada RSTAA más próxima, Bridge ID del AB asociado y vencimiento del registro. La tabla se accede por dirección IP.

La dimensión máxima de esta tabla según requisitos es de 100.000 entradas. Este requisito puede relajarse si los dispositivos diversos (sensores, etc) utilizan sistemas de resolución de direcciones específicos o servidores ARP especíalizados. Es previsible que buen número de sensores, por razones económicas, utilicen tecnologías inalámbricas específicas que reducirán dicho requisito.

En la medida en que los servidores ARP/AB implementen funciones adicionales tales como autenticación, medidas antisuplantación, servicios adicionales (DHCP u otros), el almacenamiento requerido aumentará de acuerdo con dichos requisitos, pero no se contempla ahora por no figurar en la opción preferente elegida (servidores ARP/AB solamente).

6.6 Rendimiento conjunto.

En las secciones precedentes se ha mostrado el rendimiento de los niveles de Núcleo (troncal) y de Acceso de forma separada para mayor precisión en la evaluación de los protocolos constituyentes. En la práctica el tráfico real utilizará ambos niveles de forma variable y ello dependiendo de los tipos de tráfico dominantes. Analicemos brevemente los criterios para combinar los resultados precedentes a fin de evaluar el rendimiento esperable de las redes campus propuestas.

El esquema general aplicable a los caminos utilizados en la red campus consta de uno a tres tramos: un tramo intra-área, un tramo en el troncal y un tramo final intra-área como se indica en la Figura 73. El tráfico intra-área (Figura 74) no utiliza el troncal, por lo que su rendimiento corresponde al evaluado para el nivel de Acceso. En cuanto al tráfico interárea podemos distinguir entre el tráfico interárea entre sistemas finales y el tráfico de área a servidor (tráfico cliente-servidor desde sistemas finales de un área a un servidor no situado en el área). El tráfico interárea normal utiliza caminos de tres tramos por lo que su longitud/coste medio corresponde a la suma de una mitad de camino intra área (más precisamente: un camino hasta el bridge raíz (ARB), un camino de troncal y otra mitad de camino intra-área. Combinando estas dos partes, podemos, en términos globales aproximados, considerar la suma de los caminos medios intra-área (completo) y troncal como indicador del camino medio total para el tráfico interárea.

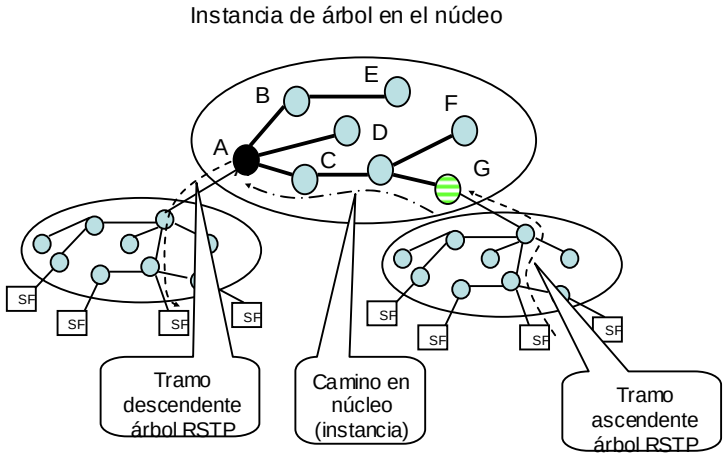


Figura 73. Caminos (mínimos) inter-área

Dado el predominio del tráfico cliente-servidor en las redes campus actuales, los servidores compartidos por las áreas estarán conectados a los ARBs del troncal de forma muy directa para optimizar el acceso a los mismos como se muestra en la Figura 74.

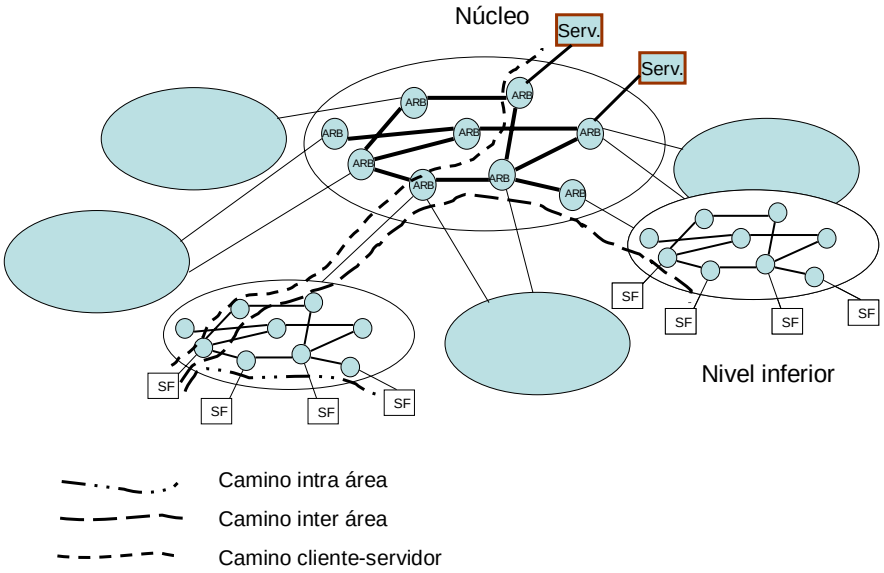


Figura 74. Tipos de caminos en red y conexión de servidores.

Esto hace que el camino de un sistema final a un servidor se componga de dos tramos: un semicamino de área desde el sistema final hasta el ARB y un tramo en el troncal entre ARBs (además del enlace entre servidor y ARB).

Por tanto una estimación de caminos medios en las redes campus debe incluir una ponderación relativa de los diferentes tráficos que la aproxime al tráfico real (a sabiendas de la impredecibilidad a medio plazo de dicha matriz de tráfico). En las redes campus actuales predomina el tráfico cliente-servidor, por lo que el rendimiento del protocolo de troncal será determinante respecto al del protocolo de Distribución-Acceso. Recordemos que todos los costes de camino de los tráficos cliente-servidor e interárea son mínimos, quedando excluidos los caminos intraárea.

En cada red concreta a evaluar deberá pues estimarse el porcentaje de tráfico intraárea respecto del total para valorar la conveniencia o no de su optimización o a efectos de comparación de soluciones alternativas, quedando descartada la optimización cuando sus valores sean pequeños, con valores tales como 20% de tráfico intraárea.

Existen aspectos importantes de costes a considerar para valorar el rendimiento conjunto. El coste de las redes campus varía mucho entre el núcleo, el nivel de distribución y el de acceso [FOLS][Wit05]. Esto se debe a las diferencias de requerimientos de prestaciones, fiabilidad, disponibilidad, características sofisticadas requeridas y a las diferencias de volúmenes de fabricación existentes. Para evaluar costes totales en la red propuesta debe ponderarse el coste de utilización/no utilización de un enlace en la red inferior respecto a uno del núcleo en términos económicos de coste de adquisición y mantenimiento. Una de las diferencias principales de coste es la diferencia en el coste por puerto de conmutador. De los 50 $ por puerto de conmutador de Acceso para Fast Ethernet a los 700-1000 $ por puerto de conmutador existe un factor de aproximadamente entre 14 y 20.

Dado el incremento de velocidades de Ethernet, los requisitos de latencia para la red se satisfacen con facilidad. Por ello la optimización de caminos debe orientarse a la mejor utilización y economía de infraestructura de red. Por otra parte, desde el punto de vista del reparto del retardo, dada la superior velocidad del núcleo, los mayores causantes de retardo son los saltos en la red inferior por su menor velocidad (diez veces menor). Lo que equivale a decir que para optimizar costes se debe actuar sobre el núcleo y para optimizar retardos sobre el acceso. Pero, como hemos visto, una vez fijada la topología, las posibles mejoras de camino solamente son posibles en los caminos intra área.

Como se ha mostrado más arriba, en el nivel de Acceso el margen de mejora máximo posible del camino medio entre árbol de expansión y conectividad total es de aproximadamente hasta un 100 % (2 saltos), al precio de emplear bridges RSJ en vez de bridges estándar, no siempre justificado. Se ha mostrado que los caminos medios optimizados son de 2,30-2,58 para núcleos de 16-32 nodos, obtenibles con AMSTP en una topología hipercúbica frente a los caminos medios de 4.0 saltos obtenibles en redes aleatorias. Una mejora de un salto en troncal supone un ahorro neto de ancho de banda a instalar computable en el coste de los puertos de conmutador de troncal economizados. Aunque el protocolo de troncal AMSTP es aplicable a redes de todo tipo, la elección de topología del troncal tiene impacto significativo en la optimización de la red.

6.7 Doble bridge raíz

Para validar la propuesta de funcionamiento de los ARB en modo activo-activo, se ha utilizado el programa *spantreesim* [STJS], modificado para aceptar empates en la elección de bridge raíz. El programa visualiza el proceso de elección de bridge raíz en STP y las BPDUs intercambiadas entre los bridges. De esta forma, si existen dos bridges (ARB) con igual prioridad para ser bridge raíz se aceptan ambos, fragmentando la red en dos subredes, cada una dependiente del correspondiente ARB. Se parte de una red en malla de 9 nodos. Los bridges *a* e *i* se configuran con mínima e igual prioridad equivalente. El estado al comenzar el protocolo de árbol de expansión se muestra en la Figura 75 (todos los bridges en rojo) y la red final tras la elección de bridge raíz muestra a los bridges *a* e *i* como raíz cada uno de una subred (Figura 76). Los segmentos de red se muestran en vertical (S1 a S12). Los trazos horizontales gruesos indican puerto seleccionado como puerto raíz por el bridge conectado.

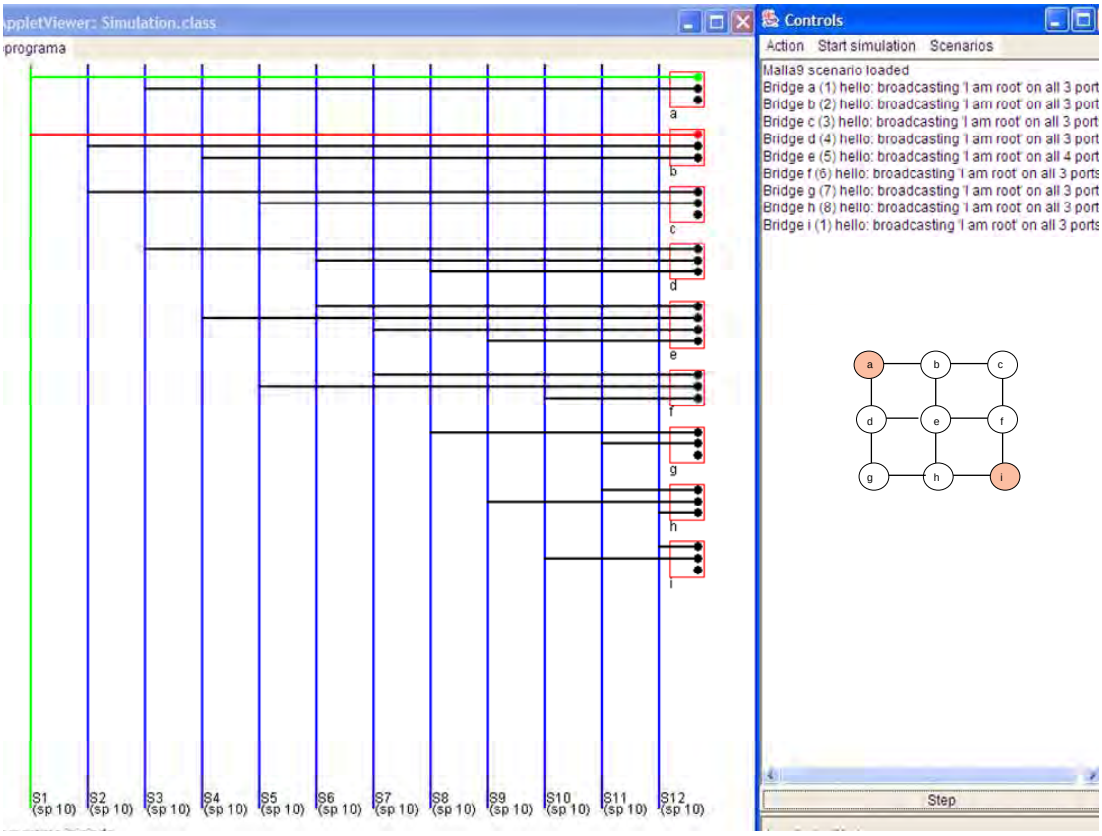


Figura 75. Disposición inicial.

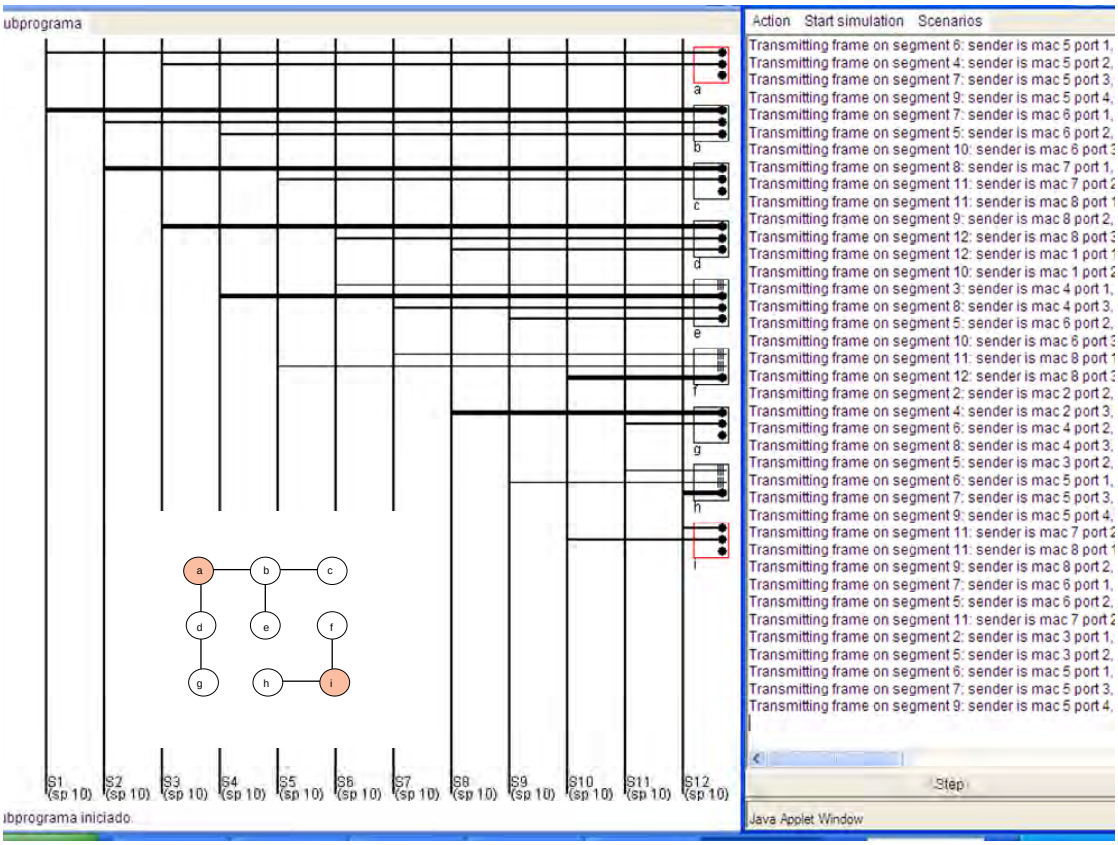


Figura 76. Disposición final de árboles de expansión.

6.8 Conclusiones

En este capítulo se ha mostrado la adecuación de la arquitectura propuesta para redes campus de tamaño medio y grande por su autoconfiguración, economía, caminos mínimos, simplicidad y uso eficiente de la infraestructura más costosa. Se mostrado la idoneidad del nuevo protocolo propuesto, AMSTP, como protocolo autoconfigurable de camino mínimo para el troncal redes campus de altas prestaciones y las prestaciones esperables. AMSTP distribuye el tráfico entre múltiples árboles de expansión por el camino más corto a destino, utilizando de forma muy eficiente la red óptica troncal, de alto coste. Los árboles múltiples permiten la utilización de todos los enlaces sin riesgo de bucles, evitando la restricción a topologías de tipo árbol o anillo, mucho menos eficientes. Se ha mostrado que AMSTP requiere una mínima configuración de los ARBs, similar a la de un bridge estándar, lo que significa economía de operación y mayor fiabilidad por evitarse los fallos debidos a de errores de configuración. Se han mostrado la baja complejidad de los mensajes, procesado simple y reducidas necesidades de almacenamiento que los protocolos propuestos, especialmente en la arquitectura simplificada, requieren en los dispositivos.

Se ha mostrado que la utilización de RSTP en la capa de Acceso, en el caso de la arquitectura simplificada, puede ser es asimismo muy eficiente en rendimiento, simplicidad y economía. Es posible una alta utilización de la red porque, debido al predominio del tráfico de y hacia el bridge raíz (por el tráfico predominante cliente-servidor), no es necesaria alta conectividad con enlaces adicionales al árbol de expansión, solamente para asegurar la fiabilidad ante fallos de enlaces. Esta arquitectura es plenamente consistente con el diseño predominante de switches Ethernet para agregación de tráfico, como los de tipo N FE + 2 * 1 GE o equivalentes de capacidad superior.

Se ha mostrado una alternativa para la redundancia de conexiones al troncal basada en configuraciones ARB activo-activo. El empleo de árboles múltiples de expansión jerarquizados en las redes campus crea nuevos escenarios de aplicación de los mismos.

Se ha mostrado que los caminos entre sistemas finales (cliente-servidor, host-router) de la red campus son mínimos en la arquitectura compuesta y en la simplificada, a excepción de los caminos dentro de una misma subred de acceso, que solamente son optimizados en la arquitectura compuesta.

Capítulo 7

Conclusiones y trabajo futuro

7.1 Conclusiones

La presente Tesis se enmarca dentro del ámbito de las redes campus Ethernet con Conmutadores de Funcionalidad Añadida, para proponer una arquitectura, dispositivos y protocolos para redes campus Ethernet de gran tamaño autoconfigurables. Las redes campus Ethernet autoconfigurables comienzan formalmente a ser objeto de atención en el IETF y en el IEEE bajo enfoques diferentes. Como consecuencia de dicho interés, en el IETF se ha creado en Julio de 2005 el grupo de trabajo Transparent Interconnection of Lots of Links (TRILL, también denominado Rbridges). El New Standard Committee (Nescom) del IEEE ha aprobado en Septiembre de 2005 el proyecto denominado Shortest Path Bridging (SPB), comenzando así el proceso de elaboración de un estándar como enmienda a IEEE 802.1Q (P802.1aq).

La primera contribución realizada en esta Tesis ha consistido en una revisión de las propuestas realizadas desde la aparición de los bridges, tanto transparentes como de encaminamiento en origen, que puedan considerarse alternativas al actual estándar de árbol de expansión y de bridges transparentes con aprendizaje. Este proceso ha incluido una revisión detallada de los estándares vigentes en el momento Spanning Tree Protocol (IEEE 802.1D), Rapid Spanning Tree protocol (IEEE 802.1w) y Multiple Spanning Tree protocol (IEEE 802.1s). Se ha revisado igualmente la evolución de las topologías y arquitecturas de redes campus y su problemática de autoconfiguración. Asimismo se han revisado las áreas relacionadas con el direccionamiento y encaminamiento en redes tales como los mecanismos usados en Peer to Peer y el encaminamiento jerárquico y sus variantes en redes fijas y móviles con vistas a su posible aplicación en la propuesta.

La segunda contribución consiste en una propuesta completa de arquitectura para redes Ethernet, basada en dos niveles jerárquicos Ethernet. Es aplicable a redes campus de gran tamaño, utiliza eficazmente la red instalada, es adecuada por su sencillez para construir conmutadores de altas prestaciones y coste moderado y requiere configuración mínima. Esta propuesta presenta mejor escalabilidad que las propuestas basadas en estado de enlaces como Rbridges y su coste es presumiblemente inferior para prestaciones equivalentes debido a la menor complejidad de proceso y almacenamiento en los conmutadores.

Las contribuciones de esta arquitectura se desglosan a su vez en las siguientes propuestas de:

- Conmutadores de Funcionalidad Añadida: Bridge de troncal de camino mínimo (ARB) para encaminamiento por árboles múltiples de expansión en el troncal y bridge de acceso (AB) para encaminamiento alternativo al árbol de expansión en la red de acceso.
- Servidores/registradores ARP/ARB en reparto de carga. Implementan un caching coordinado para la resolución de paquetes ARP y de ARB ó AB destino.
- Un protocolo denominado Protocolo Alternativo de Árbol de Expansión Múltiple (Alternative Multiple Spanning Tree Protocol, AMSTP) para los nodos del núcleo de la red campus. El protocolo requiere configuración mínima, utiliza la infraestructura de forma eficiente y distribuye el tráfico por caminos de longitud mínima entre los bridges frontera del núcleo. La idoneidad de los protocolos simplificados de árboles múltiples enraizados en los bridges extremos del camino ha sido refrendada recientemente con la propuesta inicial del IEEE de Shortest Path Bridging (Marzo 2005).
- Un Protocolo de Asignación de Coordenadas/Identificadores a bridges basadas en Árbol de Expansión Rápido (Rapid Spanning Tree Based Address Assignment (RSTAA). Este protocolo tiene otras aplicaciones potenciales, adicionales a la expuesta en esta Tesis, que utilicen identificadores de nodo sobre switches. El protocolo puede implementarse como simple extensión del protocolo estándar RSTP. La adaptación de los bridges existentes a este protocolo mediante migración software es simple.
- Un protocolo, denominado RSTAA-STAR Jerárquico (RSJ), para el nivel inferior de la red o para la red completa. El protocolo extiende el Árbol de Expansión estándar y posibilita el encaminamiento a través de enlaces cruzados entre AB, enlaces normalmente bloqueados por el protocolo de Árbol de Expansión. El protocolo es una extensión del protocolo STAR que utiliza RSTP y coordenadas jerárquicas RSTAA, de mejor escalabilidad que STAR.
- Finalmente, como un ejercicio específico de aplicación de la arquitectura propuesta para el núcleo, se han calculado varios ejemplos de red con núcleos de altas prestaciones, equivalentes a un conmutador de M*10 Gbps, construídos con conmutadores de gama media (n * 1 Gbps) interconectados con topologías de alta conectividad y se han obtenido los costes de dispositivo resultantes para equivalencia de prestaciones. Los resultados muestran un importante margen de coste respecto a los conmutadores de M*10 Gbps y un apreciable margen respecto a parejas de conmutadores centralizados de alta capacidad de tecnología N*1 Gbps.

7.1.1 Participación en grupos de estandarización.

Describimos aquí las actividades relacionadas con grupos de estandarización. En lo que respecta a participación y seguimiento de grupos de trabajo en organismos de elaboración de estándares relativos al tema objeto de esta Tesis, se han seguido las discusiones esporádicas discusiones en el IETF dentro la lista de discusión del grupo Rbridges [Rbridge] y se ha participado en los temas pertinentes a la Tesis. Se han evidenciado algunas diferencias en requerimientos (como el requisito de soportar redes muy grandes) y de aproximación de base (árboles de expansión frente a encaminamiento MAC utilizando Djikstra), por lo que no se ha sometido propuesta formal alternativa, al considerarla además fuera del ámbito IETF y dentro del ámbito de competencia del IEEE. En Julio de 2005 ha sido aprobada la constitución en el IETF del Working Group TRILL. Su definición no contempla expresamente redes campus de gran tamaño lo que separa sus objetivos de los de la presente Tesis y a su vez hace menos crítica que en nuestro problema la escalabilidad del encaminamiento basado en direcciones MAC. Se contemplan en TRILL explícitamente protocolos de encaminamiento (inicialmente IS-IS). El borrador actual de Rbridges contempla el uso de proxies ARP (como función o no del Rbridge), en combinación con ARP estándar en caso necesario, sin definirse dentro del estándar posibles optimizaciones como la propuesta de servidores/registradores ARP. El grupo TRILL mantiene un enlace con el IEEE sobre el tema, pero no se contempla actualmente trabajo conjunto entre ambos. La arquitectura de red propuesta en esta Tesis, por utilizar árboles de expansión en capa dos, nos parece más cercana al ámbito del IEEE que al del IETF, salvo en lo relativo a los servidores ARP. Finalmente, es de destacar la creciente utilización en el grupo TRILL de la terminología de los protocolos de árboles de expansión, tales como árboles de distribución por VLANs, BPDUs de cambio de topología y otras, lo que muestra una progresiva convergencia en el enfoque (pero sobre IS-IS).

En lo relativo al IEEE se han seguido las actualizaciones de los estándares de árboles de expansión 802.1D y 802.1Q. No se han producido discusiones relativas al tema de esta Tesis, que el autor conozca, hasta Abril de 2005 con la propuesta inicial de Shortest Path Bridging [Sea05], la cual ha originado una propuesta de Proyecto, aprobada en septiembre de 2005, para la elaboración del estándar 802.1aq. Se estudiará en lo sucesivo la participación en el mismo y/o la submisión de propuestas.

7.1.2 Posicionamiento relativo

En este epígrafe se sitúa a la arquitectura propuesta en esta Tesis respecto a las propuestas actuales en los grupos de estandarización en el IETF e IEEE.

Respecto a la propuesta actualmente en discusión en el IETF (WG TRILL), en cuanto a requisitos: como se ha indicado la arquitectura propuesta en la Tesis satisface el requisito de redes campus de tamaño muy grande mientras que TRILL contempla solamente los tamaños actuales de redes campus. TRILL actualmente contempla el empleo de VLANs a través del núcleo de las redes campus, VLANs que deben

configurarse, mientras que la propuesta de esta Tesis no contempla VLANs en el núcleo de la red a efectos de encaminamiento por entrar en contradicción con el requisito de autoconfiguración e incrementar la complejidad del encaminamiento. En la propuesta, el empleo de multicast con inspección IGMP es suficiente para una limitación adecuada de la difusión de multicast. En cuanto a implementación, TRILL no contempla explícitamente una capa troncal sino que los Rbridges entre sí establecen una red superpuesta que atraviesa las islas de bridges estándar intermedias. Esto mejora la compatibilidad en el núcleo con los bridges estándar, pero limita el rendimiento, la fiabilidad y la escalabilidad del mismo y aumenta la complejidad, requiriendo alterar la trama en cada salto, siendo de escasa utilidad en el núcleo, donde solamente se emplean los conmutadores de más prestaciones. En cuanto al encaminamiento, TRILL ha preseleccionado el protocolo IS-IS, modificable para operar con direcciones MAC de sistemas finales, como protocolo de encaminamiento entre Rbridges. El encaminamiento por caminos mínimos se realiza mediante árboles de expansión calculados mediante el algoritmo Djikstra a partir de los LSP recibidos de los Rbridges de forma independiente por VLANs. La propuesta de esta Tesis, basada en árboles múltiples, utiliza por tanto Vectores de Distancia a los bridges raíz, por lo que la complejidad es mucho menor, manteniendo sin embargo la convergencia rápida mediante los probados mecanismos estándar de reconfiguración de RSTP. La resolución del sistema final y del ARB destino se realiza en los servidores ARP/ARB asignados mediante hash por lo que no se sobrecargan los nodos con la función de *proxy*. Se evita asimismo el almacenamiento e intercambio entre todos los Rbridges de la red campus de las listas de sus sistemas finales conectados ni la construcción de las grandes tablas de encaminamiento resultantes.

Respecto a la propuesta inicial, del IEEE, denominada Shortest Path Bridging [Sea05]: La propuesta inicial para SPB es compatible con el aprendizaje de MACs en las VLANs, pero presenta un problema no resuelto, consistente en asegurar que los caminos de árbol establecidos en ambos sentidos coincidan para que el aprendizaje de MAC y el reenvío basado en dicho aprendizaje funcionen correctamente. Mientras que Shortest Path Bridging precisa que los caminos del árbol entre dos nodos *a* y *b* sean idénticos en el árbol con raíz en *a* y en el árbol con raiz en *b* (en dirección opuesta) para que el reenvío basado en el aprendizaje de direcciones MACs funcione correctamente (enviar a una dirección MAC por el puerto por donde han sido recibidas tramas procedentes de esa dirección), AMSTP no lo requiere porque no utiliza aprendizaje de MACs para reenviar al ARB destino, encaminando siempre por el puerto raíz del ARB origen hacia el ARB (bridge raíz de ese árbol) destino. Para que los caminos coincidan en SPB, los costes de los enlaces deben ser simétricos y solamente debe existir un camino de igual coste para evitar empates en el protocolo de formación de las instancias de árboles. Para coordinar la resolución de empates durante la formación de árboles en SPB existe una propuesta inicial (*reflection vectors)*[Fin05], no exenta de inconvenientes como la limitación de instancias de árboles, el retraso en la convergencia del protocolo por las comunicaciones adicionales necesarias de extremo a extremo del árbol y el alargamiento de las BPDUs [Sea05].

7.2 Trabajo futuro.

La existencia de sendos grupos de trabajo en IETF e IEEE relacionados con esta Tesis influye en parte en el trabajo futuro. Como se ha indicado, la arquitectura propuesta para el núcleo es más cercana a la aproximación del IEEE basada en árboles múltiples de expansión que a la del IETF. La discusión en el IETF habida previamente a la aprobación del grupo de trabajo ha madurado la solución, por lo que a priori no se vislumbran grandes cambios sobre la propuesta inicial, basada en el protocolo IS-IS para encaminamiento MAC. Por otra parte, el interés mostrado por los participantes en las votaciones parece ser moderadamente favorable en el caso del IETF y unánime en el caso del IEEE. Por estas razones, y dada la mayor importancia del núcleo en el rendimiento total, en opinión del autor, es probable una mayor proyección de los resultados de esta Tesis en el grupo del IEEE.

En el ámbito del problema estudiado, son frecuentes las soluciones basadas en paso de funcionalidad desde capa tres a capa dos, como es el caso del encaminamiento. El autor considera que la siguiente funcionalidad a incorporar en capa dos es el equilibrado de carga en las zonas de bridges Shortest Path Bridging (en el núcleo en nuestra propuesta). Aunque actualmente el estándar de agregación de enlaces [802.3ad] contribuye de forma importante a la escalabilidad y fiabilidad de las redes Ethernet, la distribución de carga se limita al ámbito de los enlaces entre dos nodos contiguos, sin funcionalidad que tenga visión de equilibrado de la carga a nivel de red. Las líneas de trabajo que se ven más prometedoras o urgentes en éste área son las siguientes:

- Aplicación del modelo desarrollado en [CVP93] para el análisis de disponibilidad de las redes propuestas y para el diseño de núcleos de alta disponibilidad.
- Estudio de introducción de mecanismos de reparto de carga en protocolo de troncal (AMSTP), basados en los conceptos de redes Valiant [ZM04] u otros. Comparación de prestaciones de troncales AMSTP frente a los basados en redes Valiant de carga equilibrada.
- Aplicación del protocolo RSTAA como auxiliar en otros protocolos tales como prevención de bucles mediante prohibición de giros, autenticación y seguridad en capa dos o identificación de nodos o enlaces.
- Prestaciones y comportamiento del protocolo RSJ en redes jerárquicas de gran tamaño.
- Análisis de compatibilidad e interoperabilidad de las soluciones propuestas para redes campus grandes con las propuestas en discusión en los organismos de estandarización para L2VPN.
- Análisis e implementación de protocolo simplificado de árboles múltiples de expansión basado en N árboles independientes (N*RSTP) descrito en el Capítulo 5, a partir del protocolo RSTP.

Definiciones

Adaptive Routing Bridge. Utilizamos el termino genérico Routing bridges para englobar a todo tipo ARB para los bridges evolucionados del troncal y el termino AB dado que no se atiene al concepto de RBridge del grupo RBridge, limitado a encaminamiento basado en direcciones MAC. Posteriormente en la propuesta distinguimos dos tipos de ARB: AB y ARBs segun su papel en la arquitectura propuesta, ambos son variantes de los ARB genericos.

Agent Bridge (Bridge Agente). Empleamos la nomenclatura de [STAR]. El bridge agente es el AB encargado de todos los sistemas finales situados debajo de él en el árbol de expansión que no tienen otro AB más cerca ascendiendo por el mismo. Si el bridge designado del AE no tiene funcionalidad AB, el primer AB hacia arriba en el AE es el bridge Agente para todos esos sistemas finales situados bajo él.

Bridge Agente RSTAR. Es el bridge RSTAR más cercano al host. Es el primer bridge RSTASTAR bridge encontrado ascendiendo por el árbol de expansión desde el sistema final.

AMSTI. Alternative Multiple Spanning Tree Instance. Cada instancia de árbol creada con el protocolo AMSTP, cuyo bridge raíz es un bridge frontera del troncal o core, utilizada para enviar por camino mínimo todas las tramas que entran en el troncal por dicho bridge frontera hacia los diferentes bridge frontera de salida del troncal

Bridge MST. Bridge capaz de soportar el CST, una o varias MSTI y tratar las tramas de forma selectiva vía CST o una determinada MSTI.

Configuration Table. Tabla de configuración MST. Tabla configurable que asigna cada VLAN al CST o una instancia MSTI.

CST. Common Spanning Tree. Árbol de expansión calculado por STP, RSTP, MSTP para conectar regiones MST.

CIST Common and Internal Spanning Tree. Árbol de expansión calculado por STP, RSTP más la conectividad a través de las regiones y bridges MSTP calculada por MSTP para asegurar la interconexión simple y completa.

Designated Bridge (Bridge Designado). El término bridge designado se emplea en el mismo sentido que en el estándar 802.1D. Es el bridge que encamina el tráfico de una LAN determinada. El bridge designado de un segmento LAN es el bridge designado para todos los sistemas finales de esa LAN.

Dual rooting. Enraizamiento dual. Red de bridges que puede tener dos bridges raíz activos, sobre cada uno de los cuales se crea un árbol de expansión constituyendo así redes separadas disjuntas, que se interrumpen en el punto de distancia equivalente a ambos bridges raíz.

Edge Bridge. Bridge frontera de la red troncal.

IST. Internal Spanning Tree. Árbol Interno de Expansión de una Región MST. Parte del CIST incluida dentro de una región MST.

Layer 3 switch. Conmutador de capa 3**.** Dispositivo que combina conmutación tipo bridge con encaminamiento en capa 3. Los conmutadores multicapa de nivel 3 encaminan las primeras tramas para un determinado destino y las siguientes las reenvían en capa 2 mediante el conmutador situado en paralelo al encaminador.

MIB. Management Information Base. Base de datos de gestión.

MSCI. Multiple Spanning Tree Configuration Identifier**.** Identificador de configuración de árboles múltiples de expansión. Nombre asignado que incluye identificador de versión para identificar una determinada asignación de VLANs a Árboles de Expansión.

MSTI. Multiple Spanning Tree Instance. Instancia de árbol multiple de Expansión.

MSTP Multiple Spanning Tree Protocol. Protocolo de Árbol Multiple de Expansión.

Multiárbol completo (*Complete Multitree*). Es el conjunto de instancias de árboles de distribución de una región, en el que cada Bridge Frontera es raíz de una instancia de árbol. El conjunto completo de las mismas interconecta todos los bridges del troncal por caminos mínimos.

Puerto de Troncal. (*Core Port)*Son los puertos de un bridge de troncal conectados a otros bridges similares del troncal, que forman parte de un bridge que ejecuta el protocolo AMSTP u otro protocolo entre bridges de troncal.

Puerto de Distribución. Denominamos *puertos de distribución* a los puertos de un bridge frontera de troncal conectados a bridges que no pertenecen al troncal sino al nivel inferior de la red. Funcionan como puertos de un bridge estándar RSTP en cuanto al protocolo de Árbol de Expansión.

Región MST. Un conjunto de LANs y bridges MST conectados físicamente mediante puertos de los bridges MST. Todos los bridges MST deben tener igual MSCI.

Región AMSTP. Es el conjunto de bridges y enlaces entre los que se ejecuta el protocolo AMSTP. Normalmente solo existe una region AMSTP por red campus.

Registro_AM. (*AM_record)* **.** Información adjunta en la BPDU correspondiente a cada instancia del árbol múltiple, conteniendo identidad del bridge raíz de dicha instancia, distancia al mismo, indicadores (flags) y otros.

Routing Bridge. Bridge Encaminador. Dispositivo genérico que combina funcionalidad de encaminador y de bridge.

RBridge: Routing Bridge con funcionalidad según la propuesta en grupo IETF RBridge [RBridge]

Sobrecarga del protocolo. *Overhead.* Ancho de banda de la red consumido por el protocolo para la comunicación entre nodos.

Tráfico Intra-área. Tráfico con originado en un área del nivel inferior y con destino en la misma área.

Tráfico Inter-área. Tráfico en la red campus originado en un área del nivel inferior (Acceso) con destino en otra área de la red campus. Atraviesa la red troncal.

Abreviaturas

10 GE. 10 Gigabit Ethernet.

AB. Adaptive Bridge.

ADSL. Asymmetric Digital Subscriber Line

AMSTI. Alternative Multiple Spanning Tree Instance.

AMSTP. Alternative Multiple Spanning Tree Protocol. Protocolo Alternativo de Árboles Múltiples de Expansión.

ARB. Adaptive Routing Bridge.

ARP. Address Resolution Protocol.

ATM. Asynchronous Transfer Mode.

CST. Common Spanning Tree.

CIST Common and Internal Spanning Tree.

DHCP. Dynamic Host Configuration Protocol.

EIGRP. Enhanced Interior Gateway Routing Protocol

ES-IS. End System to Intermediate System (protocol).

FE. Fast Ethernet.

GARP. Generic Atribute Registration Protocol.

GE. Gigabit Ethernet.

GOE. Global Open Ethernet.

GMRP. Generic Multicast Registration Protocol**.**

GVRP. Generic VLAN Registration Protocol.

HTTP. Hypertext Transfer Protocol.

ICMP. Internet Control Message Protocol.

IDF. Intermediate Distribution Frame.

IEEE. Institute of Electrical and Electronic Engineers.

IETF. Internet Engineering Task Force.

IGMP. Internet Group Management Protocol.

IS-IS. Intermediate System to Intermediate System.

IST. Internal Spanning Tree.

LLDP. Link Layer Discovery Protocol.

MDF. Main Distribution Frame..

MIB. Management Information Base.

MPLS. Multiprotocol Label Switching.

MSCI. Multiple Spanning Tree Configuration Identifier**.**

MSTI. Multiple Spanning Tree Instance.

MSTP Multiple Spanning Tree Protocol.

NAT. Network Access Translator.

NDP. Neighbour Discovery Protocol.

PAE. Protocolo de Árbol de Expansión..

P2P. Peer- to-Peer**.**

RSJ. Protocolo RSTAA-STAR Jerárquico

RSTAR. RSTP-based STAR.

RSTP. Rapid Spanning Tree Protocol IEEE 802.1D 2004.

SAN. Storage Area Network.

SEND. Secure Neighbour Discovery Protocol.

SDH. Synchronous Digital Hierarchy.

SPB. Shortest Path Bridging (IEEE).

STAR. Spanning Tree Alternate Routing Protocol

STP. Spanning Tree Protocol IEEE 802.1D 1998.

VDSL. Very high rate Digital Suscriber Line.

VLAN. Virtual LAN.

Apéndice A Modelo RSJ

El protocolo RSJ es una extensión del protocolo STAR [STAR] que hace uso de las nuevas coordenadas de tipo jerárquico RSTAA para realizar el encaminamiento. En este apéndice se describe el modelo formal de RSJ. Este modelo es una extensión y adaptación del modelo del protocolo STAR [Lui02] al uso de dichas coordenadas.

La red de bridges se representa mediante un grafo *G = (R, D, E).* En dicho grafo *R* es el conjunto de bridges RSJ , *D* es el conjunto de bridges estándar 802.1D y *E* es el conjunto de enlaces que inteconecta los bridges. Representamos un enlace entre los bridges *a* y *b* como como (*a*,*b*).

Llamamos *c(x, y)* al coste del enlace (x, y). Se suponen costes no negativos. Se asigna coste infinito a los enlaces que no pertenecen a *E*. Cuando existe más de un enlace conectando a dos bridges determinados, se utiliza el coste menor de ellos entre los que estén disponibles.

Definimos un camino en el grafo *G* como una secuencia de enlaces pertenecientes a *E* y que no tiene bucles. El coste del camino es la suma de los costes de los enlaces del camino. La distancia entre un par de nodos es la longitud del camino más corto que une los nodos.

- **Definiciones**

Vecino directo. El bridge x es *vecino directo* de *y* si $(x, y) \in E$.
$A = (V, E_A)$ es un subgrafo en árbol del grafo *G* que representa un Árbol de Expansión (AE), en el que $(x, y) \in E_A$ si y solo si *(x, y)* es un enlace activo del AE.
Enlace de árbol. Enlaces de árbol son los que pertenecen a E_A y enlaces *extra (*no-árbol) son los que pertenecen a $E - E_A$
Vecinos de árbol. Si $(x, y) \in E_A$, *x e y* son *vecinos de árbol.*
Camino de árbol. Un camino de *A* que une un bridge origen *o* con un bridge destino *d* es un *camino del árbol,* que se especifica como *camino_de_árbol(o, d).*

Cuando los bridges x e y son vecinos de árbol, la distancia cumple que $d_A(x, y) = c(x, y)$. La distancia del camino de árbol desde *x* al bridge raíz $d_T(r; x)$, se denomina distancia del camino raíz de *x*. Nos referimos a un *camino_de_árbol(o; d)* como árbol de bridge estándar si tiene al menos un bridge intermedio, excluyendo a los bridges origen y destino y todos los bridges intermedios del camino son estándar. Si *R* es el conjunto de bridges RSJ. Si *o* y *d* son bridges RSJ , es decir *o* y $d \in R$ y existe al menos un camino de árbol bridge estándar entre ellos, *o* es *vecino distante* RSJ y viceversa.

Tomando como referencia el bridge raíz, podemos asimismo definir relaciones *padre-hijo* y *antecesor-descendiente* entre los bridges. Sear *r* el bridge raíz. Un bridge *a* es antecesor de otro bridge *b* si el *camino_de_árbol(b; r)* pasa por *a*. Dicho de otra forma, *b* es *descendiente* de *a*. Definimos como nivel de un bridge a el número de saltos del *camino_de_árbol(a; r).* El nivel es superior cuanto menor es el número de saltos desde el bridge raíz.

El *antecesor común más cercano acmc(a,b)* de *a* y *b* es el bridge de mayor nivel del camino de árbol entre *a* y *b*, de otra forma, el bridge intersección de menor nivel de los caminos de árbol (a; r) y (b; r).

Si x es un *antecesor* de y, entonces x es necesariamente el *antecesor común más cercano* de x e y.

Decimos que x e y están en *distinta rama* si $acmc(x,y) \neq / x$ y también $\neq$ y. Decimos que $(x,y) \in E\text{-}E_T$ es un *enlace cruzado* si *x* e *y* están en distinta rama. En este caso *x* e *y* son *vecinos* distantes.

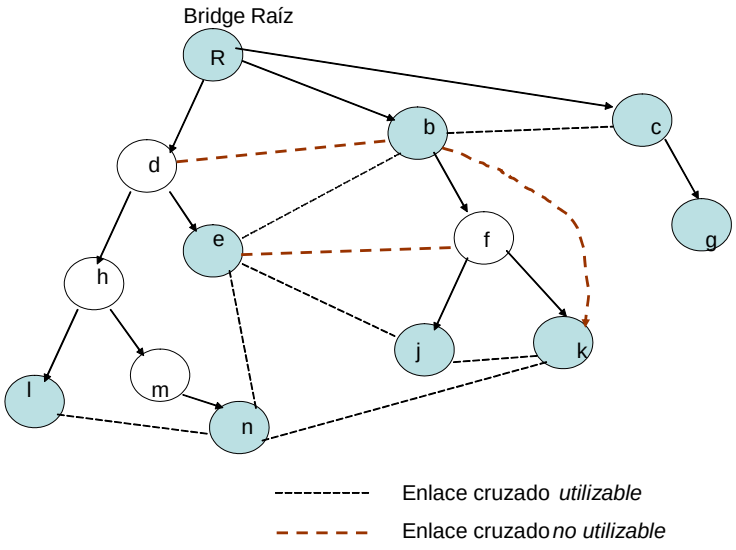


Figura 77. Enlaces cruzados en RSJ

Tabla 10. Tipos de vecinos

Tipos de vecinos

Tipo de vecino	Definición	Ejemplos
Vecinos directos	*x,y € V, (x,y)€ E*	*R-d, b-d*
Vecinos de árbol	*x,y € V, (x,y)€* E_T	*b-f, f-k*
Vecinos de enlace cruzado	*x,y € V, (x,y) €* E-E_T y acmc(x,y)=/ x e y	*b-k, e-b, d-b*
Vecinos directos RSJ	*x,y € B, (x,y)€ E*	*e-b, e-n, n-k*
Vecinos distantes RSJ	*x,y € B, camino_árbol (x,y) es un camino de bridges 802.1D*	*b-k, r-e*

En la Figura 78 se muestra un grafo no dirigido de una LAN con bridges. El nodo R es el bridge raíz. Los nodos blancos son nodos estándar y los sombreados son nodos RSJ. Se muestran con línea continua los enlaces del árbol de expansión y con linea discontinua los enlaces que no pertenecen a dicho árbol. Hay que distinguir entre enlaces que no pertenecen al árbol y enlaces cruzados. En la figura, *b-k* es un enlace que no pertenece al árbol pero no es utilizable como enlace cruzado. Se distinguen en la figura porque unos se representan con linea discontinua de punto y los otros con rayas. En la Figura 78 se muestran las definiciones de vecinos y ejemplos de los mismos de forma sumaria. Los bridges *b* y *k* son *vecinos RSJ distantes* porque en el camino del árbol existe un bridge estándar. El enlace *b-k* no es utilizable porque no aporta camino más corto que el de el árbol de expansión (si fuera más corto habría sido elegido por el protocolo de AE). Llamamos pues *enlaces cruzados utilizables* a los que pueden ser utilizados como caminos más cortos que el árbol de expansión. Solamente los bridges *vecinos RSJ* pueden utilizar el enlace que les une como enlace cruzado, ya que los bridges estándar bloquean los puertos de los enlaces no pertenecientes al árbol de expansión.

A fin de facilitar la discusión del proceso de búsqueda de caminos más cortos alternativos entre bridges RSJ, definimos el grafo superpuesto (overlay) de un grafo normal con bridges. Este grafo consta solamente de bridges RSJ. Lo definimos como el grafo RSJ de un grafo *G = (B, D, E)* y lo denominamos G_B. Existe enlace entre dos bridges RSJ siempre que sean *vecinos RSJ distantes* o *vecinos RSJ directos* en el grafo G.

Definición. *Grafo de bridges RSJ. Un grafo RSJ* G_B *está formado exclusivamente por bridges RSJ. Formalmente,* $G_B = (B, E_B)$ *donde B es el conjunto de bridges RSJ y (x,y)* $\in$ E_B *si y solamente si x e y son vecinos RSJ, distantes o directos. c'(x,y) es el coste del enlace (x,y) en* E_B *y está definido por:*

c'(x,y)= c(x,y) si x e y son vecinos directosRSJ pero no distantes

= $min (d_T(x,y), c(x,y))$ *si x e y son tanto vecinos directos como distantes.*

=$d_T(x,y)$ *si x e y son vecinos distantes pero no directos*

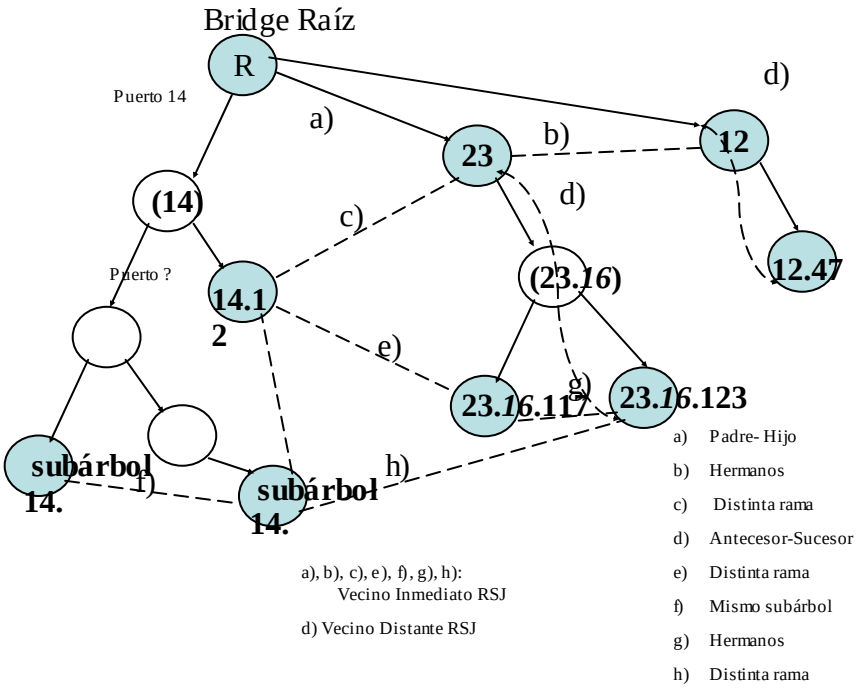


Figura 78. Relación topológica entre bridges RSJ

El grafo RSJ de la red de Figura 78 se muestra en la Figura 79. El coste de cada enlace se muestra junto al enlace. La métrica considerada es número de saltos

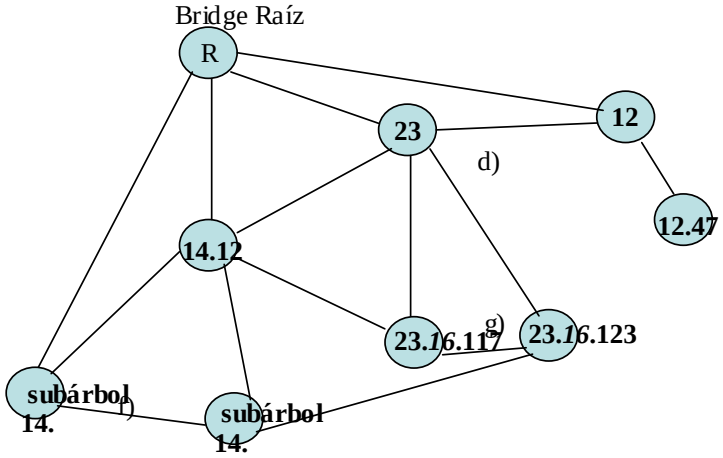


Figura 79. Grafo de bridges RSJ

Apéndice B Redes *scale free*

En este apéndice hacemos una breve referencia a las redes *scale free,* cuyo relativamente reciente descubrimiento ha desplazado algo el foco de investigación desde las propiedades de las redes, dirigiendo la atención a sus mecanismos de formación y crecimiento. En [AB02] se describen las propiedades y evolución de las redes desde las redes random tradicionales y en [Bar02] se describe la historia de los conceptos *small-world* y *scale-free* de forma cualitativa.

Albert-Laszlo Barabasi y otros investigadores de la Universidad de Notre Dame, estudiando en 1998 los enlaces entre páginas web, que esperaban de características propias de redes aleatorias, descubrieron que tenían propiedades similares a los fractales. Por ello acuñaron el término *scale free* para las redes en las que el grado de conectividad es altamente irregular y a la vez persistente en todo el rango de valores. A diferencia de las redes aleatorias, en las que el grado de los nodos se distribuye alrededor y cerca del valor medio, la distribución del grado de los nodos en estas redes sigue una recta si se representa en escala doble logarítmica, de acuerdo a una ley de potencia inversa del grado de los nodos. Aunque el número de nodos de grado *d* en la red decae fuertemente al aumentar *n*, existen con cierta probabilidad nodos con grado elevado y para cualquier grado. En el caso de las páginas web, la probabilidad de que un nodo con muchos enlaces obtenga nuevos enlaces es proporcional a su número de enlaces. En estas redes los nodos de alto grado actúan como conectores (*hubs*) para un gran número de nodos de baja conectividad. A su vez los nodos de conectividad superior respectivamente actúan como conectores de conectores, de ahí el término que alude a la inexistencia de escala. La comparación más gráfica entre redes aleatorias en sentido tradicional y scale free utiliza una red de carreteras como ejemplo del primer caso (probabilidad de camino ligada a distancia entre nodos) y un mapa de rutas aéreas como ejemplo de las scale free. Las redes scale free presentan características de Small-World, pero no a la inversa necesariamente.

El fenómeno se ha descrito también como las leyes de potencia inversa de Internet (Power Laws at Internet). [FFF99][AB02]. La ecuación que expresa la distribución de conectividad en los nodos es la siguiente:

$$n(d) \approx \frac{K}{d^{\alpha}}$$

donde n es el número de nodos de grado d en la red, K es una constante y α tiene un valor entre 2 y 3. Dicho de otro modo, unos pocos nodos acaparan la alta conectividad. Esta alta conectividad indirectamente reduce el número de saltos, produciendo el efecto mencionado de acortamiento de caminos.

Independientemente del rango de grados de nodos que se estudie, la proporción entre los de grado alto y bajo se mantiene, dicho de otra forma, la pendiente de la curva de distribución de grados es constante si se representa en un gráfico doble logarítmico.

El descubrimiento de esta ley ha influido el estudio de las redes complejas en búsqueda de aspectos comunes [Cnr05] [Str01]. La aplicabilidad de esta ley ha sido estudiada en muy diversos tipos de redes, incluida la conectividad de los encaminadores en Internet. No tenemos noticias de estudios al respecto en el ámbito de las redes campus. Esto es debido probablemente a varias razones, siendo la primera el reducido tamaño relativo de las mismas si se consideran de forma separada de las redes externas. Otra razón es la transparencia e "invisibilidad" relativa de los bridges en las redes actuales, que hace estas redes más difíciles de estudiar (normalmente solamente personal autorizado puede tener conocimiento detallado de la topología de la red campus). Otra razón es probablemente el tamaño reducido de dichas redes hasta fechas recientes, lo que no las ha hecho dignas de estudio, situación que está cambiando con la extensión de Ethernet a redes metropolitanas. Por otra parte el diseño ex profeso de estas redes bajo una autoridad única y bajo criterios de cierta jerarquización para facilitar su gestión hace que las tipologías se reduzcan y dependan sobre todo de la tecnología utilizada y la fecha en que se introdujo. Las redes campus se diseñan con unos criterios determinados y con unos tipos de switches tipificados de variedad muy limitada, lo que reduce bastante la variabilidad de la conectividad, haciendo menos probable su cumplimiento. En todo caso el estudio de las topologías reales de las redes campus aportaría escenarios de comparación y validación de las nuevas propuestas, más alla de escenarios teóricos de redes de bridges interconectados de forma aleatoria con costes variables.

El efecto de acortamiento de camino producido por los nodos "concentradores", no es exclusivo de las redes *scale free*, y fue puesto de manifiesto mucho antes por el famoso estudio de Milgram de *small World effect* [Mil67] se produce, independientemente de las características *scale free*, por la mera existencia de nodos de alta conectividad en la red.

Apéndice C Diseño económico de núcleos.

Se analiza en este apéndice una metodología de diseño que utiliza algunos de las propuestas de esta Tesis en combinación con el ciclo de vida de las tecnologías Ethernet y se desarrollan algunos ejemplos. Los precios de los conmutadores Ethernet descienden de forma continuada a medida que la producción de la nueva tecnología alcanza volumen masivo. Esto ha ocurrido con las sucesivas versiones de Ethernet 10 Mbps, 100 Mbps y 1 Gigabit. Los dispositivos de mayor consumo como las tarjetas controladoras de red (NIC) y los conmutadores de sobremesa, bajan sus precios más rápidamente que los conmutadores más sofisticados. La evolución de los precios de los equipos está muy ligada al desarrollo de los ASICs que éstos utilizan, lo que explica el descenso más rápido del precio de los dispositivos de uso masivo. Actualmente la tecnología Gigabit Ethernet está entrando en la madurez en cuanto a los dispositivos de mayor consumo [Pricegrabber], pero aún no lo ha hecho en los conmutadores de prestaciones más exigentes. En 10 GE, sin embargo, los conmutadores han bajado de precio hasta una quinta parte de su precio en 2002, pero su precio por puerto es todavía muy elevado respecto al de los conmutadores de 1 Gbps.

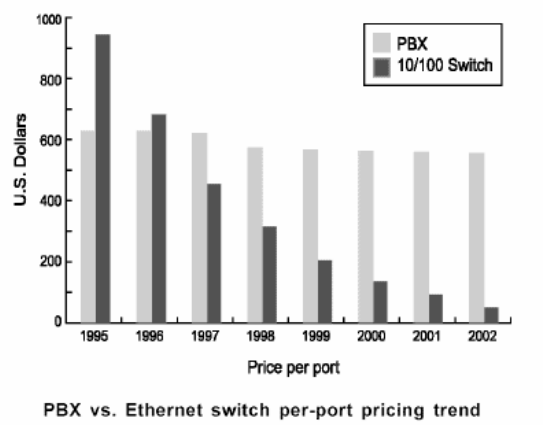


Figura 80. Evolución de costes por puerto de conmutadores Ethernet (fuente: Telos systems).

La estrategia para un diseño de alta capacidad y económico que se expone aquí consiste en utilizar la arquitectura basada en AMSTP para implementar núcleos de alta capacidad de conmutación mediante conmutadores de capacidad media-baja de tecnología madura Ethernet, actualmente utilizando conmutadores de n*1 Gbps[3]. Un núcleo basado en conmutadores fuertemente interconectados puede presentar frente a los conmutadores de núcleo centralizados ventajas económicas derivadas de su menores requerimientos así como fiabilidad equivalente utilizando componentes de inferior fiabilidad.

El coste por enlace de un conmutador 10 GE es de unos 25.000 dólares por enlace (suma de los costes por puerto de 10 Gb del conmutador core (15.000) y del de Acceso o Distribución (10.000) [Wit05]. Para 1 Gbps es de unos 1.400 $ de los que 640 corresponden a los dos GBIC.

Núcleos de topologías hipercúbicas.

En esta sección exploramos los rangos de dimensionamiento alcanzables con la configuración descrita. Se consideran tamaños de núcleos de 8 a 64 nodos con diferente número de puertos de 1 Gbps. Se detalla el caso de núcleo de 16 nodos y se compara con las alternativas para evaluar económicamente sus posibilidades. En la Figura 82 se muestran las capacidades de la topología de cubo (ocho nodos, hipercubo de grado 3) según el número de puertos de 1 Gbps por nodo. Se dimensionan de forma equilibrada los puertos de servidor respecto a los de usuarios. Los índices de sobresuscripción son de 8, 5 y 17.

La Figura 81 ilustra el caso de una red campus con núcleo de 16 nodos que utiliza 32 enlaces por nodo.

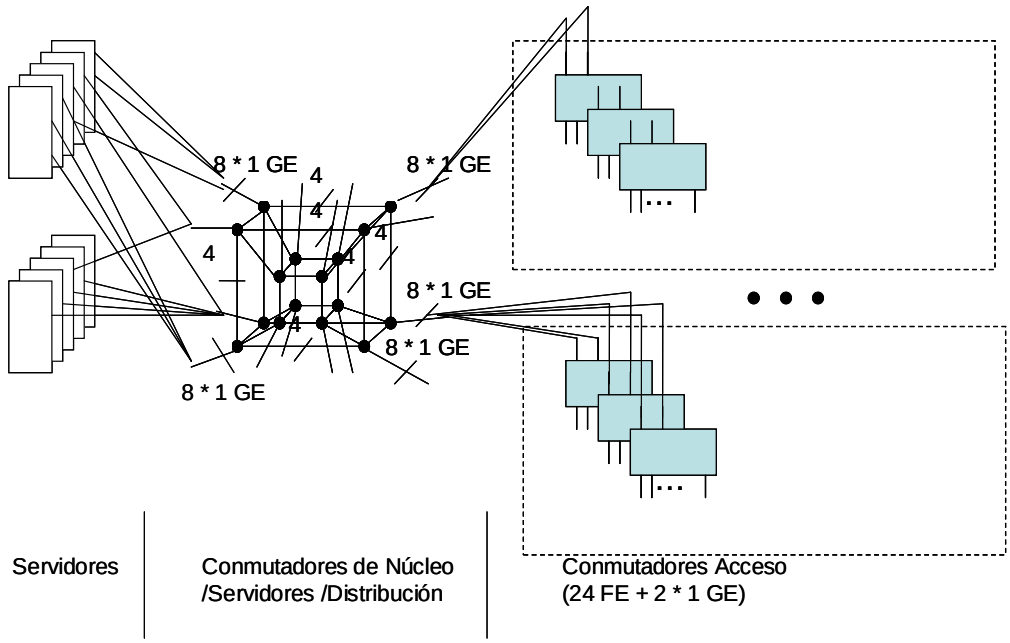


Figura 81. Red campus con núcleo de 16 nodos.

[3] Actualmente Gigabit Ethernet no ha alcanzado todavía el volumen masivo óptimo ni por ello precios óptimos para conmutadores de núcleo, pero es la mejor opción disponible. En el futuro, cuando 10 GE alcance la madurez y 40 GE sea la tecnología emergente Ethernet, se utilizarían conmutadores 10 GE para mantener la ventaja económica.

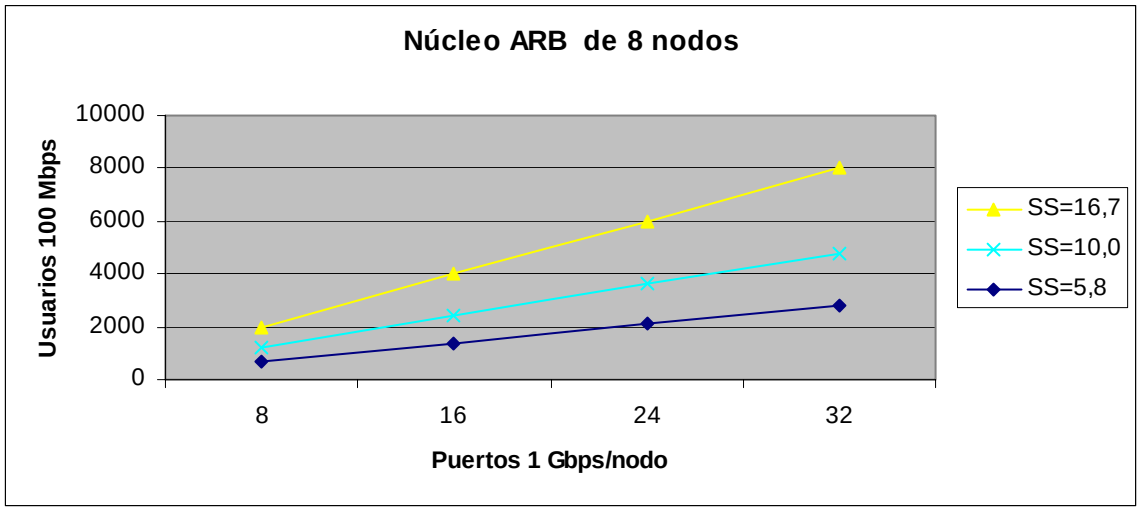


Figura 82. Capacidades núcleo de 8 nodos

Ejemplo de red 16 * 32:

El nivel troncal está formado por un núcleo de 32 nodos. Cada ARB tiene 32 puertos de 1 Gbps. Suponemos tráfico 100% cliente-servidor para el dimensionamiento de servidores. En el nivel inferior se suponen conmutadores de grupo de trabajo de 24*100 Mbps + 2*1 GE. Con la topología de doble hipercubo el grado de interconexión del núcleo para 32 nodos es de 5. Utilizando los datos de saturación de tráfico obtenidos para la topología hipercúbica de 32 nodos, el tráfico máximo ofrecido es de 2,34 (ver Tabla 6). Esto significa que si cada ARB está conectado a otros cinco ARB mediante enlaces de 1 GE, el tráfico máximo inyectable en cada ARB es de 2,34 * 0,8 = 1,872 Gbps. Por lo tanto los enlaces de cada nodo deben repartirse en esa proporción entre puertos de distribución y puertos de troncal si consideramos topología sin bloqueo. Los puertos de distribución se distribuyen entre puertos de usuarios y puertos de servidores. Realizando este dimensionamiento para otros números de puertos por nodo (8, 16 y 24 nodos) se pueden explorar las capacidades de esta topología. En la Figura 83 y en la Tabla 11 se muestran las capacidades obtenidas de la topología de 16 nodos (hipercubo de grado 4) según el número de enlaces de 1 Gbps por nodo. Los índices de sobresuscripción para los casos calculados son de 6.2, 10.3 y 17.4.

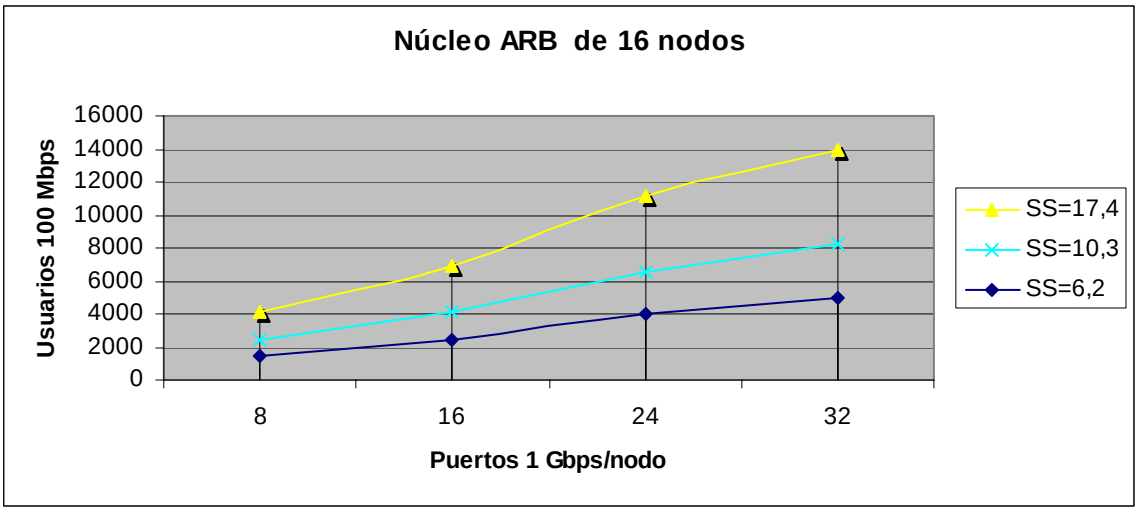


Figura 83. Características topología núcleo de 16 nodos.

Tabla 11. Características núcleo de 16 nodos

Características núcleo ARBs 16 nodos						
puertos 1 GE/nodo			8	16	24	32
num. Nodos ARB			16	16	16	16
capac. Bruta total núcleo			128	256	384	512
puertos servidor 1 GE req.			24	40	64	80
tráfico saturación l topologia 16 nodos 2,34 *0,8			1,872	1,872	1,872	1,872
puertos de distribución por ARB			3	5	8	10
puertos totales distribución			48	80	128	160
puertos usuarios			24	40	64	80
usuarios		SS=17,4	4180	6950	11150	13900
usuarios		SS=10,3	2460	4100	6600	8240
usuarios		SS=6,2	1480	2460	3950	4950

A continuación se comparan los costes de tres opciones para un núcleo de 160 Gbps de capacidad total, equivalente al obtenible en la configuración máxima de 16 nodos. La primera opción utiliza un conmutador de núcleo de 16 puertos de 10 Gbps [Wit05]. La segunda utiliza un conmutador de núcleo equipado con 160 puertos de 1 Gbps, en dos chassis de 80 puertos cada uno [FOLS]. La tercera opción utiliza 16 conmutadores ARB de 32 puertos de 1 Gbps cada uno. En la Tabla 12 se muestra el resultado. La opción basada en 10 GE es muy costosa dada la inmadurez de la tecnología. El precio por ARB que se muestra es el precio límite por ARB que igualaría el precio con la segunda opción.

Tabla 12. Comparación de costes de núcleos.

		Precio unitario	Total
Precio 16*10 GE Alcatel 8800 opt 2			
Base			50995
Puertos	16	9995	159920
Opt. Transceiver	16	4995	79920
			581670
Precio 160 * 1 GE (FOLS)			
			20407
Blades 16 puertos 1000 FX	5	8015	40075
GBICs	160	321	51360
			111842
	2		**223684**
Precios núcleo 16 nodos * 32 enlaces/nodo			
(para equivalencia a 160*1 GE)			
Conmutadores ARB (precio límite)	16	**10770**	172324
GBICs	160	321	51360
			223684

En la Figura 84 se muestran en iguales condiciones las capacidades de la topología de 32 nodos (hipercubo de grado 5) según el número de puertos de 1 Gbps por nodo.

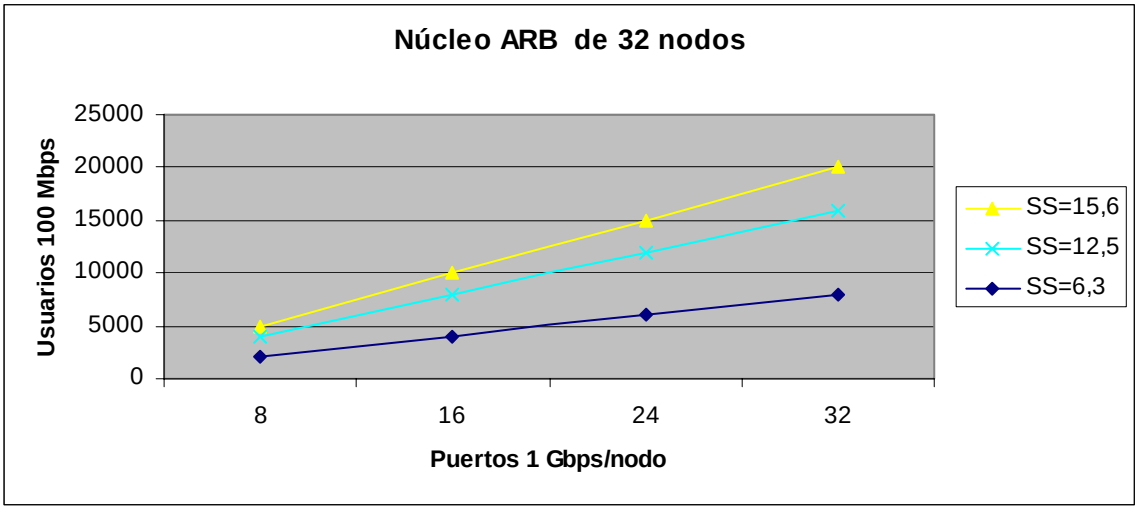


Figura 84. Características núcleo topología 32 nodos.

En la Figura 85 se muestran en iguales condiciones las capacidades de la topología de 64 nodos (hipercubo de grado 6) según el número de enlaces de 1 Gbps por nodo.

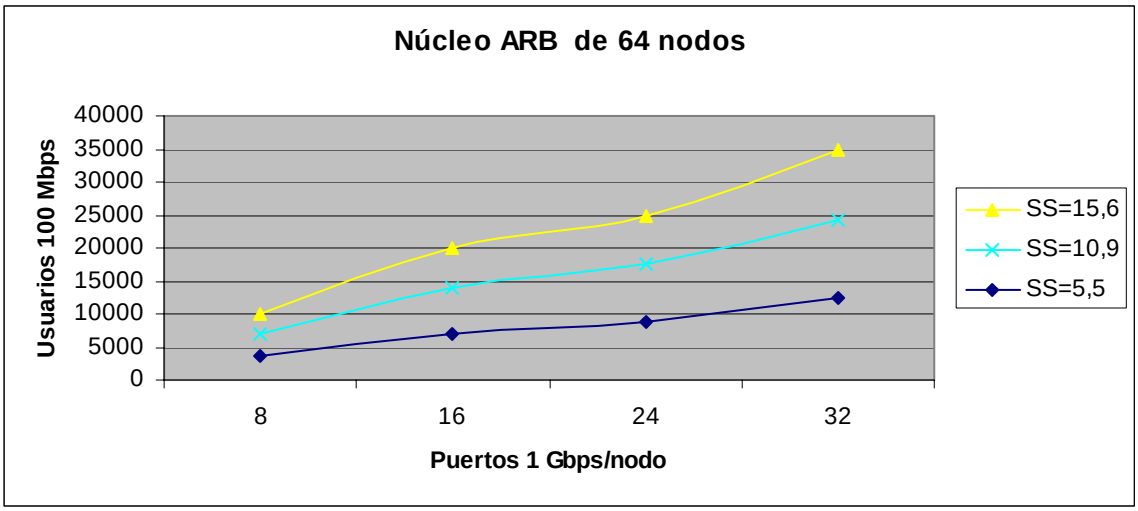


Figura 85. Características núcleo de 64 nodos

En la medida en que los accesos de usuarios a Internet aumentan su ancho de banda con el incremento de capacidad de ADSL de los abonados, el ancho de banda requerido en las redes corporativas aumentará igualmente, al popularizarse de forma creciente el acceso a contenidos multimedia on-line. Por ello es presumible que las variables de diseño de redes campus evolucionen hacia consumos de importantes ancho de banda a medida que el coste de dicho ancho de banda se equipare al de los equipos actuales. La denominada regla 2X/10X postula, para predecir la adopción masiva de cada nueva tecnología Ethernet, que los usuarios migran a la tecnología Ethernet de capacidad 10 veces superior cuando su coste es del orden de dos veces superior al de la tecnología Ethernet existente.

En cuanto a la ubicación de los ARBs una vez eliminada la capa de distribución como tal al unirla al núcleo, puede ser la de los conmutadores de distribución, normalmente colocados en la planta baja de cada edificio, o la de conmutadores de núcleo. Normalmente parece recomendable ubicarlos como núcleo por razones de mantenimiento, seguridad y normalmente también de costes, para minimizar las interconexiones de fibra óptica.

Apéndice D Núcleo de conectividad total

En este apéndice se muestran las prestaciones obtenibles con núcleos formados por ARBs conectados en topologías de conectividad total.

Las topologías de conectividad total entre nodos suponen la máxima capacidad de tráfico obtenible y la vez el máximo coste de enlaces y puertos, dado el máximo grado de conectividad de estas redes (N-1). Si se utilizan árboles múltiples de expansión, todos los caminos son de un solo salto. El ejemplo considerado para la comparación es un núcleo de 8 nodos con conectividad total, como se muestra en la Figura 87 y en la Figura 86.

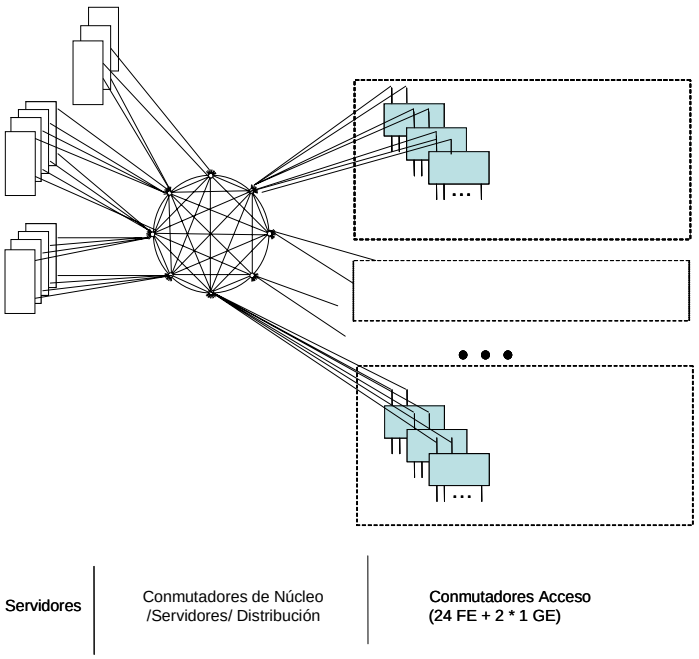


Figura 86. Red campus con núcleo de 8 nodos de conectividad total.

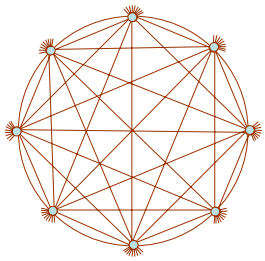


Figura 87. Núcleo de 8 nodos con conectividad total (9+7 enlaces/nodo).

El diseño y las prestaciones se muestran en la Figura 88 y la Tabla 13. Se han considerado los casos de 8 ARBs con 16, 32 y 48 puertos de 1 Gbps por ARB. La red no presenta bloqueo si se emplean la mitad de los puertos de cada ARB para puertos de usuarios y la otra mitad para puertos de troncal. Dado que los enlaces entre nodos del troncal son 7 o multiplos de 7, de cada 16 puertos se asignan 9 puertos como puertos de usuarios y 7 para troncal, por lo que existe una probabilidad de bloqueo aunque reducida.

Tabla 13. Características núcleo ARB 8 nodos conectividad total.

Características núcleo ARBs 8 nodos conectividad completa				
puertos 1 GE/nodo		16	32	48
num. Nodos ARB		8	8	8
puertos de distribución por ARB		9	18	27
puertos totales distribución		72	144	216
puertos usuarios		36	72	108
usuarios	SS=17,4	6280	12500	18750
usuarios	SS=10,3	3700	7400	11100
usuarios	SS=6,2	2230	4440	6650

La capacidad de esta red de ocho nodos es aproximadamente un 50% superior a la de la topología cúbica del mismo número de nodos. La mejora obtenida se debe al número de enlaces (28-12 = 16) adicionales.

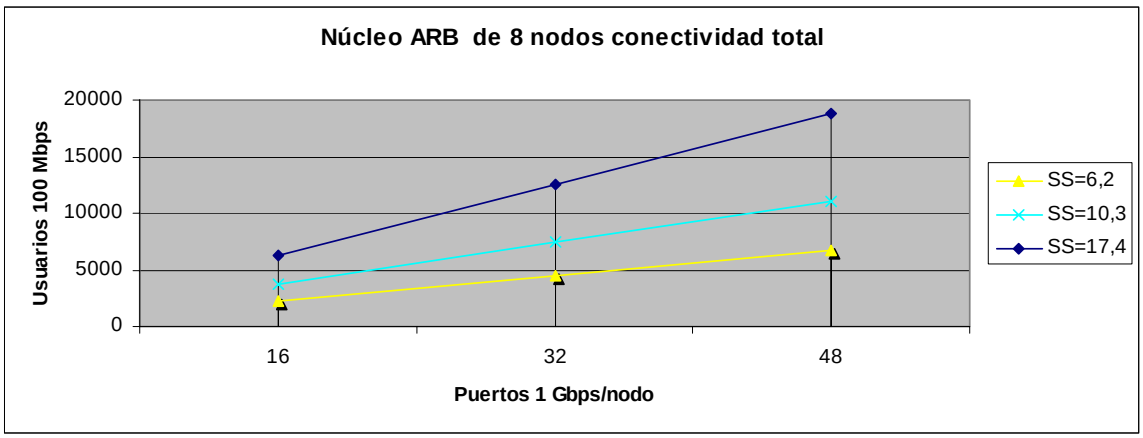


Figura 88. Características núcleo ARB 8 nodos conectividad total.

El inconveniente que presenta esta topología es la insuficiente granularidad de la capacidad del núcleo en cuanto al número de enlaces por nodo, de un enlace de 1 Gbps

entre cada nodo del núcleo se pasa a dos, tres, etc. El número de puertos troncales consumidos es pues: N-1, 2*(N-1), 3*(N-1). Este número debe ser cercano a la mitad (o proporción fija) de los del conmutador ARB, por lo tanto la mitad del número de enlaces por nodo debe ser múltiplo de N-1 (o ligeramente superior para no desperdiciar enlaces).

Una característica de interés de estas topologías es la facilidad para implementar mecanismos de distribución de carga. y prevenir la congestión, dada su máxima conectividad.

Referencias

[802.1ad] *P802.1ad™. Draft Standard for Local and Metropolitan Area Networks-- Virtual Bridged Local Area Networks-- Amendment 4: Provider Bridges.* IEEE, Enero 2005.

[802.1D04] *IEEE 802.1D-2004 IEEE standard for local and metropolitan area networks- Media access control (MAC) Bridges.* IEEE, 2004. Disponible en [Internet]: <http://standards. ieee.org/getieee802/802.1.html > (Julio 2005)

[802.1Q] *IEEE 802.1Q-2003 IEEE standard for local and metropolitan area networks - Virtual Bridged Local Area Networks*- IEEE, 2003. Disponible en [Internet]: <http://standards. ieee.org/getieee802/download/802.1Q-2003.pdf > (Julio 2005).

[802.1s] IEEE. *802.1s-Multiple spanning trees.* Disponible en [Internet]: http://www. ieee802.org/1/ pages/802.1s. html

[802.1X] *IEEE 802.1X- IEEE standard for local and metropolitan area networks. Port based network access control.* IEEE. Disponible en [Internet]: http://www.ieee802. org/1/ pages/802.1x.html.

[802.3] *802.3™. IEEE Standard for Information technology— Telecommunications and information exchange between systems—Local and metropolitan area networks— Specific requirements Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications.* IEEE, Marzo 2002.

[AB02] R. Albert and A.-L. Barabasi. *Statistical mechanics of complex networks.* Rev. Mod. Phys. no. 74, p. 47-97, 2002.

[AFE01] Yehuda Afek et al. *Routing with a Clue*. (Tel Aviv univ). En IEEE/ACM transactions on networking vol.9 no 6. Dic. 2001.

[AI04] A. Azcorra, G. Ibáñez . *Adaptive Routing Bridges: Nuevos enfoques para redes campus conmutadas*. Mundo Internet. Madrid. Feb. 2004.

[AJY00] C. Alaettinoglu, V. Jacobson, and H. Yu. *Towards milli-second IGP convergence.* IETF draft-alaettinoglu-ISIS-convergence-00, Noviembre 2000. Disponible en [Internet]: http://www.packetdesign.com/news/industry-publications/ drafts/ convergence.pdf.

[Alt+04] H. Altunbasak et al. *Addressing the Weak Link Between Layer 2 and Layer 3 in the Internet Architecture*. IEE Proceedings of the LCN Nov. 2004, p. 417-418.

[AM03] O. Artemjev, V. Myasnyankin. *Are those loops on your network neck secure?* Disponible en [Internet]: http://olli.digger.org.ru/STP/STP-article.en.pdf (Julio 2005).

[Anr98] The Anritsu Company. *The must-have reference for multilayer switching*. Disponible en: http://www.zurich.ibm.com/pdf/ AnritsuGlossary.pdf.

[AVM02] Oleg K. Artemjev, Vladislav V. Myasnyankin. *Fun with the Spanning Tree Protocol.* Disponible en [Internet]: http://www.phrack.org/show.php?p=61&a=12 (Julio 2005)

[BA99] A.-L. Barabási and R. Albert. *Emergence of Scaling in Random Networks.* Science, p. 509–512, Oct. 1999.

[Bar02] A.-L. Barabasi. *Linked: The New Science of Networks.* Perseus, Cambridge, MA 2002.

[BCF+94] N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawlk, C. L. S. ad J. N. Seizovic, and W. Su, *Myrinet: A Gigabit-per-second local area network.* IEEE Micro, vol. 15, p. 29–36, Febrero 1994.

[Ben88] E. Benhamou. *Integrating Bridges and Routers in a Large Internetwork*. IEEE Network Magazine, 2(1), Enero 1988.

[BGM+00] Y. Breitbar, M. Garofalakis, C. Martin et al. *Topology Discovery in Heterogenous IP Networks*. IEEE INFOCOM 2000. p.265-274.

[BH88] L. Bosack and C. Hedrick. *Problems in Large LANs*. IEEE Network Magazine, 2(1), Enero 1988.

[Bla04] P.E. Black. *Dictionary of Algorithms and Data Structures. Steiner Tree*. Ed. NIST. Disponible en [Internet]: http://www.nist.gov/dads/HTML/steinertree.html. (Julio 2005)

[Bra89a] R. Braden. *Requirements for Internet Hosts -- Communication Layers. Internet Engineering Task Force. Request for Comments: 1122.* Oct. 1989. Disponible en [Internet]: http://www.ietf.org/rfc/rfc1122.txt.

[Bra89b] R. Braden. *Requirements for Internet Hosts – Application and Support. Internet Engineering Task Force. Request for Comments: 1123.* Oct. 1989. Disponible en [Internet]: http://www.ietf.org/rfc/rfc1123.txt.

[BRITE] *Boston University Representative Internet Topology Generator*. Disponible en [Internet]: http://www.cs.bu.edu/brite/ (Julio 2005)

[CAM03] *Gigabit Campus networks Design*. WhitePaper. Cisco Systems. 1999/2003. Disponible en [Internet]: <http://www.cisco.com > (Julio 2005).

[CAS+94] S. Casale, V. Catania, A. Puliafito, L. Vita. *A Remote Bridging Technique to Increase Performability in Distributed Systems*. En *IEEE Transactions on Industrial Electronics, Vol. 41*, p. 461-472. Abril 1994.

[CGMP] *Cisco Group Management Protocol*. Disponible en [Internet]: http://www. cisco.com/warp/ public/473/22.html (Julio 2005).

[CH01] Kennedy Clark, Kevin Hamilton . *Cisco LAN Switching* (CCIE Professional Development series) ISBN: 1-57870-094-9. Cisco Press. 2001.

[CHA01] C.-S. Chang, D.-S. Lee, Y.-S. Jou. *Load balanced Birkhoff-von Neumann switches, part I: one-stage buffering.* En IEEE HPSR '01, Dallas, Mayo 2001.

[CHI+02] G. Chirovolu et al. *Encapsulation schemes to extend Ethernet to Metropolitan Area Networks.* Alcatel Telecommunications Review, 3rd Quarter 2002.

[CHI+04] G. Chirovolu et al. *Issues and Approachex on Extndign Ethernet Beyond LANs.* IEEE Communications Magazine, p. 80-86, March 2004.

[Cia] R. Ciampa. *Layer 3 switching basics*. Disponible en [Internet]: http://www. pulsewan.com/data101/pdfs/layer3_switching.pdf (Julio 2005).

[CIS04]. Cisco Enterprise Marketing. *Hierarchical Campus Design*.. Disponible [Internet]:http://www.cisco.com/application/pdf/en/us/guest/netsol/ns24/c643/ cdccont0900/aecd800d8129.pdf (Mayo 2005).

[Cisco] Cisco. http:// www.cisco.com.

[CM02] B. Chen and R. Morris. *L+: Scalable landmark routing and address lookup for multi-hop wireless net-works.* 2002. MIT LCS Technical Report 837. Marzo, 2002.

[Cnr05] *Center for Complex Network Research*. Notre Dame University. Disponible en [Internet]: http://www.nd.edu/~networks/publications.htm#anchor-allpub0001.

[CPV93] V.Catania, A. Puliafito, L.Vita. *A Modular Network Architecture for Performance Enhancement in Extended Local Area Network.* IEEE Transactions on Reliability Vol 42, no. 1, March 1993.

[CZM00] Chamlee, M.A., Zegura, E.W., Mankin, A. *Design and Evaluation of a Protocol for Automated Hierarchical Address Assignment*. En Proceedings Ninth International Conference on Computer Communications and Networks, p. 328-333, 2000.

[DA01] R. Droms, W. Arbaugh. *Autentication for DHCP messages*. Internet Engineering Task Force. Request for Comments: 3118. Junio 2001. Disponible [Internet]: http://www.ietf.org/rfc/rfc3118.txt.

[Dee89] S. Deering. *Host Extensions for IP Multicasting*. Internet Engineering Task Force. Request for Comments: 1112. Agosto 1989. Disponible [Internet]: http://www.ietf.org/rfc/rfc1112.txt

[Dju97] S. Djumin. Gigabit networking: High-Speed Switching and Routing. Disponible [Internet]: http://www.cse.ohio-state.edu/~jain/cis788-97/ftp/gigabit_nets.pdf (Julio 2005).

[Dro97] R. Droms. Dynamic host configuration protocol. Internet Engineering Task Force. Request for Comments: 2131, Marzo 1997.

[DR01] P. Druschel, A. Rowstron. *Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems*. En Proceedings of the 18th IFIP/ACM Interna-tional Conference on Distributed Systems Platforms (Middle-ware 2001)W (Nov 2001).

[DP88] Roy C. Dixon, Daniell A. Pitt. *Addressing, bridging and source routing*. IEEE Network. Enero 1988.

[DS80] Edsger W. Dijkstra, C.S. Scholten. *Termination detection for diffusing computations*. Information Processing Letters, 11(1): 1-4, Agosto 1980.

[DSR04] *The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)*. Disponible [Internet]: http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt

[DYN97] J. Duato, S. Yalmachili, L. M. Ni, *Interconnection networks: An Engineering Approach.* Los Alamitos, California: IEEE Computer Society, 1997.

[EFK03] J.Eriksson, M. Faloutsos, S.Krishnamurthy. *Peernet: Pushing Peer-to-Peer Down the Stack*. En International Peer-To-Peer Symposium, IPTPS 2003 Berkeley, Feb 2003.

[FE04] M. Fiddler G. Einho. *Routing in Turn-Prohibition Based Feed-Forward Networks*. En NETWORKING 2004, Lecture Notes in Computer Science, LNCS 3042, p.1168 –1179, 2004.

[Fin05] N. Finn. Shortest Path Bridging. Julio 2005. Disponible en [Internet]: www. ieee802.org/802_tutorials/july05/nfinn-shortest-path-bridging.pdf

[FFF99] C. Faloutsos, P. Faloutsos, M. Faloutsos. *On Power-Law Relationships of the Internet Topology*. En Proceedings of the ACM SIGCOMM 1999, p. 251-262, Sept. 1999.

[FOLS05] TIA Fiber Optic LAN section. *Cost model*. Disponible en Internet: http://www.fols.org/resources/costmodel_reg.cfm. Junio 2005,

[Freenet] *Freenetworks*. www.freenet.org

[GDS03] R. García, J. Duato, F. Silla. LSOM: A Link State Protocol Over MAC Addresses for Metropolitan Backbones Using Optical Ethernet Switches. Proceedings Second IEEE NCA'03.

[GHS83] R. Gallagher, P. Humblet, P. Spira. A Distributed Algorithm for Minimum-Weight spanning trees. ACM Transactions Programming LAN and Systems. Jan 1983, p. 66-77.

[Giant] Cisco. *Understanding Baby/Giant Jumbo Frames support on Catalyst.* Disponible en [Internet]: http://www.cisco.com/warp/public/473/175.html#topic1

[GT-ITM] *GT Internetwork Topology Models (GT-ITM)*. Georgia Tech College of Computing, 1996. Disponible [Internet]. http://www.cc.gatech.edu/projects/gtitm/gt-itm (Julio 2005).

[Gnutella] *Gnutella web site*. Disponible [Internet]: http:// www.gnutella.com.

[Groove] Disponible [Internet] : http://www.groove.net

[HAD01] I. Hadzic. *Hierarchical MAC address Space in Public Ethernet Networks*, IEEE Globecom, 2001. p. 1563-1569.

[HAR89] J. Hart. *Distributed Load Sharing*. U.S. Patent 4,811,337. March 1989. http://www. freepatentsonline. com/4811337.html

[HIP05] *IETF Host Identity Protocol (HIP) Working Group page*. Disponible en [Internet]: <www.ietf.org/html.charters/hip-charter.html> , Julio 2005

[HK97] Y. Huang P. K. McKinley. *Group Leader Election under Link-State Routing*. Computer Communications, Volume 23, Issue 7, 13 March 2000, Pages 653-666.

[HOC01] P. Hochmuth. *IP: Big protocol on campus*. Disponible en : www.networkworld.com/news /2001/0521colleges.htm.

[HOR95] R. W. Horst. *TNet: A reliable system area network.* IEEE Micro, vol. 15, p. 37–45, February 1995.

[HT03] C. Hristea, F. Tobagi. *Optimizing Mobility Support in Large Switched LANs*. En Proceedings of the IEEE ICC, Anchorage, Alaska, Mayo 2003

[HXG02] X. Hong, K. Xu, M. Gerla.- *Scalable Routing Protocols for Mobile Ad Hoc Networks*. IEEE Network July/Aug 2002 Vol. p. 11-20.

[IA04] G. Ibáñez, A. Azcorra. *Application of Rapid Spanning Tree Protocol for Automatic Hierarchical Address Assignment to Bridges*. 11th International Telecommunication Networks Strategy and Planning Symposium. Networks 2004. Wien. June 2004. Disponible en [Internet]: www.ieee.org/ieee.explore (Diciembre 2004).

[IANA05] Internet Assigned Numbers Authority. *Internet Multicast Addresses.* Disponible en [Internet]: http://www.iana.org/assignments/multicast-addresses (Diciembre 2004).

[Iba03a] G. Ibáñez. *Aspectos de Seguridad en Redes Locales e Inalámbricas : Acceso a la Red controlado por puerto (IEEE 802.1X).* Congreso Iberoamericano de Telemática: CITA2003. Montevideo, Sept. 2003.

[Iba03b] G. Ibáñez. *Peer to Peer routing and addressing: Application of P2P concepts for mobile and fixed networks*. Univ. Carlos III de Madrid. 2003. Disponible en [Internet]: http://enjambre.it.uc3m.es/%7egibanez/

[IGA04] G. Ibáñez, A. García, A. Azcorra. *Alternative Multiple Spanning Tree Protocol (AMSTP) for Optical Ethernet Backbones*. En Proceedings of IEEE HSLN (LCN 2004) Tampa, Nov. 2004. Disponible en [Internet]: www.ieee.org/ieee.explore (Diciembre 2004).

[IGMP] Zyxel. *IGMP snooping.* Disponible [Internet]: http://www.zyxel.com/support/ supportnote/ ies1000/ app/igmpsnoop.htm. (Junio 2005).

[Ish04+] K. Ishizu et al. *APG-Report: SSTP: An 802.1s Extension to Support Scalable Spanning Tree for Mobile Metropolitan Area Network*. En Proceedings of Globecom 2004. Diciembre 2004. Disponible en [Internet]: <http://www2.nict.go.jp/mt/b192/ pub/ papers/globecom2004-CRL-APG-Report7.pdf> (Marzo 2005)

[ISO03] ISO. *ISO 9542 End System to Intermediate System Routing Information Exchange Protocol for Use in Conjunction with the Protocol for Providing the Connectionless-Mode Network Service*.

[ISO84] International Organization for standarization. *ISO 8473 Protocol for Providing the OSI Connectionless-Mode Network Service.*

[Iwa+04] A. Iwata et al. Global Open Ethernet Architecture for a Cost-Effective Scalable VPN Solution. En IEICE Trans. On Communications, E87-B, 1, p.142-151, Enero 2004.

[Iwa+99] A. Iwata et al. *Scalable Routing Strategies for Ad Hoc Wireless Networks*. IEEE Journal on Selected Areas in Communications, Número especial Ad-Hoc Networks, Agosto 1999, p.1369-79. Disponible en [Internet]: http://www.cs.ucla. edu/NRL/ wireless/PAPER/ jsac99.ps.gz

[JAI92] *A comparison of hashing schemes for address lookup in Computer Networks*. R. Jain. IEEE Transactions on Communications vol 40 no 3 pp 1570-1573. Oct. 1992

[JM03] J.M. Jouanigot, O. Martin. *CERN Technology Tracking for LHC. Networking Technology*. IT Division CERN. Febrero 2003.

[Kap+01] J. Kaplan et al. *Extended Ethernet Frame Size Support.* Disponible [Internet] : http://www.ietf.org/proceedings/01aug/I-D/draft-ietf-isis-ext-eth-01.txt

[Kem86] M.F. Kempf. *Bridge Circuit for Interconnecting Networks*. U.S. Patent 4.597.078. Junio 1986.

[Kes+03] I. Keslassy et al. *Scaling Internet Routers Using Optics*. En Computer Communication Review, Vol. 33, No. 4, p. 189-200, Octubre 2003.

[KK74] L.Kleinrock and F. Kamoun, *Hierarchical routing for large networks*. Computer Networks, 1977, p. 155-174.

[Kru56] J. B. Kruskal. *On the shortest spanning subtree of a graph and the travelling salesman problem.* En Proceedings Am. Math. Soc. 7(1) p. 48-50, Febreo 1956.

[KS01] G. Kuo K.C. Shu. *Design of global hierarchical routing architecture on future IPv6 Internet*. En Proceedings of Global Telecommunications Conference, 2001. GLOBECOM '01. Nov. 2001. Vol. 1. p. 121 - 125 .

[L2VPN] *Layer 2 Virtual Private Networks. Internet Engineering Task Force*. Disponible en [Internet]: http://www.ietf.org/html.charters/l2vpn-charter.html

[LG91] Y.-D. Lin and M. Gerla. *Brouter: The Transparent Bridge with Shortest Path in Interconnected LANs.* En Proceedings of LCN, 1991, p.175-183.

[LLC98] H. Lin, S. Lai, P. Chen. *An Algorithm for Automatic Topology Discovery of IP Networks*. IEEE 1998

[LLDP] *Draft Standard for Local and Metropolitan Networks: Station and Media Access Control Connectivity Discovery, IEEE P802.1AB/D13*. IEEE. Diciembre 2004.

[LLN02] K.-S. Lui, W. C. Lee, and K. Nahrstedt, *STAR: A transparent spanning tree bridge protocol with alternate routing.* En ACM SIGCOMM Computer Communications Review, vol. 32, Julio 2002.

[LLN03] K.S. Lui, W.C. Lee, K. Nahrstedt. *Link Layer Multi-priority frame forwarding*. IEEE International Conference on Communications. May 2003, pp 1573-1577 vol.3.

[LOG01] B. Lowekamp, D. R. O'Hallaron, T. R. Gross. *Topology Discovery for Large Ethernet Networks.* En Proceedings of ACM SIGCOMM 2001, Agosto 2001.

[Lui02] K.S. Lui. *Alternate Routing Protocols for Bridged Networks*. Ph.D. These. University of Urbana-Campaign. 2002.

[Lyn96] Nancy A. Lynch. *Distributed Algorithms*. Morgan Kaurmann Publishers, Inc. San Francisco, 1996.

[Mar03a] G. Marro. *Vulnerabilities on some Data Link Layer Protocols.* SecLab. Univ. Davis. Disponible en [Internet]: http://seclab.cs.ucdavis.edu/seminars/STP.ppt (Julio 2005)

[Mar03b].G. Marro. *Attacks at the Data Link Layer*. Master Thesis Davis Univ. California. Disponible en [Internet]: http://seclab.cs.ucdavis.edu/papers/Marro _masters_thesis. pdf. (Julio 2005)

[MES04] *MESH IEEE 802.11s Working Group*. Disponible en [Internet] <http://standards.ieee.org/board/nes/ projects/ 802-11s.pdf>. (Diciembre 2004).

[MEZ04] A. Myers, T.S. Eugene, H. Zhang. *Rethinking the Service Model: Scaling Ethernet to a Million Nodes*. En Proceedings of HOTNETS III. Nov. 2004

[Mil67] S. Milgram. *The small world problem*. Psychology Today 2, p. 60–67 (1967).

[MIC03] *Microsoft Windows Server 2003 TCP/IP implementation details*. Junio 2003. Disponible en [Internet]: <http://www.microsoft.com/technet/prodtechnol/windows server2003/technologies/networking/tcpip03.mspx> (Junio 2005)

[MLS] Techfest. *Multilayer switching*. Disponible en [Internet]: <http://www.techfest. com/ networking/mlayer. htm>. (Junio 2005).

[MM00] A. J. McAuley, K. Manousakis. *Self-configuring networks*. En 21st Century Military Communications Conference Proceedings, vol. 1, p.315--319, MILCOM 2000.

[MM02] P. Mainmounkov, D. Mazières. Kademlia: *A Peer-to-Peer Information Sustem Based on the XOR Metric*. En Proceedings of IPTPS 2002 (International Workshop on Peer-to-Peer Systems) 2002. Lecture Notes in Computer Science LNCS2429, pp 34-44.

[Moo65] Wikipedia. *Moore's law*. Disponible en [Internet]: http://en.wikipedia.org/wiki/ Moore's_ law

[MRR80] J.Mcquillan, I. Richier, E. Rosen. *The new routing algorithm for the ARPANET*. IEEE Trans. On Communications. Vol. 28 núm. 5, pp 711-719. 1980

[MRP] IEEE. *"Standard for Local and Metropolitan Area Networks Virtual Bridged Local Area Networks - Amendment 07: Multiple Registration Protocol"*. Disponible en [Internet]: < http://www.ieee802.org/1/pages/802.1ak.html> Julio 2005.

[MZ96] B. McDonald , T. Znati. *Comparative analysis of neighbor greeting procotols ARP versus ES-IS.* en Proceedings of IEEE 29th Annual Simulation Symposium, Abril 1996.

[NS2] *The network simulator ns2*. Disponible en [internet] <http://www.isi.edu/nsnam/ ns/ > Julio 2005.

[NSTO] *The network simulator ns-2: Topology generation*. Disponible en [Internet]: http://www.isi.edu/nsnam/ns/ns-topogen.html, Julio 2005

[OSP03] P. Pillay-Esnault, *OSPF refresh and flooding reduction in stable topologies*. IETF draft-pillay-esnault-ospf-flooding-07.txt, June 2003.

[PAR05] *Preliminary draft of PAR for Shortest Path Bridging for comment*. Disponible en : http://www.ieee802.org/1/files/public/docs2005/new-seaman-shortest-path-par-0405-02.htm.

[PE05] R.Perlman, D. Eastlake. *RBridges: Transparent Routing* http://www.ieee802.org/802_tutorials/march05/11-05-0241-00-tttt-RBridges-transparent-routing.pdf

[PEE01] *Peer-to-Peer: Harnessing the Benefits of a Disruptive Technology*. A. Oram Editor. Ed. O'Reilly, Sebastopol, CA 2001

[Pel+04] F. D. Pellegrini, D. Starobinski, M. G. Karpovsky, and L. B. Levitin, *Scalable cycle-breaking algorithms for gigabit Ethernet backbones.* En Proceedings of IEEE Infocom 2004, Marzo 2004

[Per00] R. Perlman. *Interconnections. Bridges, Routers, Switches, and Internetworking Protocols*. Second Edition. Addison-Wesley. 2000.

[Per04] R. Perlman. *Rbridges: Transparent routing.* En Proceedings of IEEE Infocom 2004, Marzo 2004.

[Per85] R. Perlman. *An Algorithm for Distributed Computation of a Spanning Tree in an Extended LAN.* En Proceedings of Ninth ACM Data Communications Symposium, Vol. 20, No. 7, p. 44–52, Septiembre 1985, New York, USA.

[PGH00] G. Pei, M. Gerla, X. Hong. *Lanmar: Landmark routing for large scale wireless ad hoc networks with group mobility*. En Proceedings ACM MobiHOC'00, 2000.

[PHL92] R. Perlman, W. Hawe, A. Lauck. *Utilization of Redundant Links in Bridged Networks.* U.S.Patent 5150360. Sept. 22, 1992.

[PHV88] R. Perlman, A. Harvey, and G. Varghese. *Choosing the Appropriate ISO Layer for LAN Interconnection*. IEEE Network Magazine, 2(1), January 1988.

[Plu82] D. C. Plummer. *An Ethernet address resolution protocol*. *Internet Engineering Task Force. Request for Comments: 826.* Noviembre 1982.

[Pri57] R. C. Prim. *Shortest connection networks and some generalisations*. Bell Systems Technical journal. p.1389-1410, Noviembre 1957.

[PriceGrabber05] PriceGrabber.com. Disponible en Internet: http://www.PriceGrabber.com.

[PW03] R. Perlman, A. Williams. *Design for a Routing Bridge. IETF draft-perlman-zerouter-cbridge-00.txt.* Junio 2003.

[PRR97] C. Greg. Plaxton, Rajmohan Rajamaran, Andrea W. Richa. *Accessing Nearby Copies of Replicated Objects in a Distributed Environment.* ACM Symposium on Parallel Algorithms and Architectures, 1997.

[PW87] D. A. Pitt, J. Winler. *Table Free Bridging*. En Proceedings IEEE JSAC SAC-5, 9. Diciembre 1987.

[PTY04] R. Perlman, J. Touch, A. Yegin. *RBridges: Transparent Routing draft-perlman-RBridge-00.* Disponible en Internet: http://www.ietf.org/internet-drafts/draft-perlman-RBridge-00.txt. April 2004.

[RBridge] *The RBridge archives*. Disponible en [Internet]: http://www.postel.org/ pipermail/ RBridge/.

[Rex+04] J. Rexford et al. *Network-wide decision making: Toward a wafer-thin control plane*," En Proceeedings of HotNets-III, Noviembre 2004.

[Rat+01] S. Ratnasamy et al. *A scalable content-addressable network.* En Proceedings of the ACM SIGCOMM, San Diego CA, Agosto 2001.

[RF91] B. Rajagopalan, M. Faiman. *Load sharing and shortest path routing in Transparently Interconnected LANs*. En Proceedings INFOCOM, 1991.

[RPR] IEEE*. IEEE Standard for Information technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific requirements- Part 17: Resilient packet ring (RPR) access method and physical layer specifications.* Disponible en [Internet]: http:/www.ieee802.org/ (Mayo 2005).

[RSS02] S. Ratnasamy, S. Shenker, and I. Stoica. *Routing algorithms for DHTs: Some open questions.* En IPTPS, 2002.

[RSTP] *LAN/MAN Standards Committee of the IEEE Computer Society, IEEE Standard for Local and metropolitan area networks–Common Specifications Part 3: Media Access Control (MAC) Bridges–Ammendment 2: Rapid Reconfiguration*, Junio 2001. Disponible [Internet]: http://www.ieee802.org/ 1/pages/ 802.1w. html

[RTA00] T. L. Rodeheffer, C. A. Thekkath, and D. C. Anderson, *SmartBridge: A scalable bridge architecture*. En Proceedings of ACM SIGCOMM 2000, Agosto 2000.

[SC88] D. Sincoskie and C. Cotton. *Extended Bridge Algorithms for Large Networks*. IEEE Network Magazine, 2(1), Enero 1988.

[SD97] F. Silla and J. Duato. *On the Use of Virtual Channels in Networks of Workstations with Irregular Topology*. En Proceedings of the 1997 Parallel Computing, Routing, and Communication Workshop, Junio 1997. Lecture Notes In Computer Science; Vol. 1417. Proceedings of the Second International Workshop on Parallel Computer Routing and Communication p. 203 - 216 ISBN:3-540-64571. 1997.

[Sea99] M. Seaman. *Loop cutting in the original and rapid spanning tree algorithms*. IEEE. Noviembre 1999. Disponible en [Internet]: <http://www.ieee802.org/1/files/ public/docs999/loop_cutting08.pdf/ >

[Sea05] M. Seaman. *Shortest Path Bridging*. Disponible en: http://www.ieee802.org/1/ files/public/docs2005/new-seaman-shortest-path-par-0405-02.htm.

[Sei00] R. Seifert. *The Switch Book: The complete guide to LAN switching technology*. John Wiley & Sons, 2000.

[Sei88] William M. Seifert. *Bridges and Routers*. IEEE Network. Enero 1988.

[SGF02] Rüdiger Schollmeier, Ingo Gruber, Michael Finkenzeller. *Routing in Mobile Ad-hoc and Peer-to-Peer Networks.* Networking 2002 Workshops, LNCS 2376, p. 172-186, 2002.

[Sha+04] S. Sharma, K. Gopalan, S. Nanda, and T. Chiueh. *Viking: A multi-spanning-tree Ethernet architecture for metropolitan area and cluster networks*. En Proceedings of IEEE Infocom 2004, Marzo 2004.

[Sho+91] M. Shoreder et al.. *Autonet: A High-Speed, Self-Configuring Local Area Network Using Point-to-Point Links*. IEEE Journal on Selected Areas in Communications, Vol. 9, No. 8, p. 1318-1335, October 1991.

[SKZ02] L. Starobinski, M.G. Karpovsky, L. Zakrevski.-*Application of Network Calculus to General Topologies using Turn-Prohibition*. IEEE INFOCOM 2002 p. 1151-1159. 0-7803-7476-2/02.

[SM00] L.H. Sahasrabuddhe, B. Mukherjee. *Multicast Routing Algorithms and Protocols: A Tutorial*. IEEE Network, En./Feb. 2000, p. 90-102.

[Sto+00] I. Stoica et al. *Chord: A Scalable Peer-to-Peer Lookup Protocol for Internet Applications.* En Proceedings SIGCOMM 2001. Agosto 2001. p. 149-160.

[Sto+02] I. Stoica et al. *Internet indirection infrastructure*. En Proceedings ACM SIGCOMM 2002, Agosto 2002.

[STJS] E. Lesage. *Spanning Tree Java Simulator*. Disponible en [Internet]: www.iro.umontreal.ca/ lesagee/spantreesim/

[STP] *IEEE 802.1D.IEEE standard for local and metropolitan area networks--Common specifications--Media access control (MAC) Bridges*. 1998.

[Str01] S. H. Strogatz. *Exploring Complex Networks*. Nature, 2001. Disponible en [Internet]: http://www.nature.com
(Julio 2005)

[TG91] T.-Y. Tai, M. Gerla. *LAN Interconnection: A Transparent, Shortest-Path Approach.* En Proceedings ICC '91, 1991. p. 1666-1670.

[Tou+03] J. Touch, Y. Wang, L. Eggert, G. Finn, *Virtual Internet Architecture.* Future Developments of Network Architectures (FDNA). En Proceedings of SIGCOMM. Agosto 2003. ISI-TR-2003-570.

[Tsu88] P.Tsuchiya. *The Landmark Hierarchy: a new hierarchy for routing in very large networks.* En Proceedings on Communications architectures and protocols 1988. p. 35 - 42.

[VB81] L. Valiant, G. Brebner. *Universal schemes for parallel communication.* En Proc. of the 13th Annual Symposium on Theory of Computing, p. 263-277, May 1981.

[Vernon] R. D. Vernon. *Network Service Billing Strategies at Cornell*. Disponible en [Internet]: http://www.cit.cornell.edu/oit/Arch-Init/netbilling/ (Julio 2005)

[VLAN] *IEEE 802.1s/D15. Draft standards for Local and Metropolitan Area Networks- Virtual Bridged Local Area Networks- Amendment 3 to 802.1Q virtual bridged local area networks: Multiple Spanning trees.*

[VPLS] *Virtual Private LAN Service*. Disponible en [Internet]. http://www.vpls.org. (Agosto 2005).

[WH00] B.Wang, J.Hou. *Multicast Routing and Its QoS Extension: Problems, Algorithms, and Protocols.* IEEE Network, January- February 2000, p. 22-36.

[Wit05] Art Wittman. *RFP: Next-Generation Ethernet Switching.* Disponible en: www.networkmagazine.com/shared/article/showArticle.jhtml?articleId=52600272.

[Yu+03] Heeyeol Yu, Shirshanka Das, Yujin Lim, Mario Gerla. *Efficient Building Method of Multiple Spanning Tree for QoS and Load Balancing*. En Proceedings Globecom, IEEE 2003 Global Communications Conference. Diciembre 2003, p. 3620-3625.

[ZB89] Magda El Zarki, Saewoong Bahk. *A Routing Scheme for Hierarchically structured LAN/WAN Networks*. (Univ. Pennsylvania). En Proceedings MILCOM ´89, 1989 .
º

[ZCB96] Ellen W. Zegura, Ken Calvert and S. Bhattacharjee. *How to Model an Internetwork.* En Proceedings of IEEE Infocom '96, San Francisco, CA

[ZFB04] L. Zier; W. Fischer; F. Brockners; *Ethernet-based Public Communication Services : Challenge and Opportunity.* IEEE Communications Magazine, vol. p. 88-95, March 2004.

[Zha+02] B. Y. Zhao, et al. *Brocade: Landmark Routing on Overlay Networks*. IPTPS 2002 (International Workshop on Peer-to-Peer Systems) 2002. En Lecture Notes in Computer Science. LNCS2429, pp 34-44.

[ZKJ01] B.Y Zhao, J. Kubiatowicz, A. Joseph. *Tapestry: An infrastructure for fault-tolerant wide-area location and routing.* Tech. Rep. UCB/CSD-01-1141, University of California at Berkeley, Computer Science Department, 2001.

[ZM04] R. ZhangShen, N. McKeown. *Designing a Predictable Internet Backbone Network.* En Proceedings of Hotnets III. San Diego, Nov. 2004.